An investigation into the possibilities and limitations of decoding heard, imagined and spoken phonemes using a low-density, mobile EEG headset

Scott Wellington

Master of Science
Speech and Language Processing
School of Philosophy, Psychology & Language Sciences
University of Edinburgh

2019

Abstract

In this research we investigate the constraints and affordances of using a low-density EEG headset for the purpose of decoding heard, spoken and imagined phonemes into audio using the EEG signal data alone. As part of this investigation, we also introduce a new, publicly-available dataset that we compiled for this purpose[1], satisfying a database 'gap' that exists for researchers pursuing similar investigations.

We explore several preprocessing techniques for both the EEG and the audio data to achieve our regression-based decoding objective, and propose a novel stacked decoder architecture to improve decoding accuracy. We conclude that it is possible—using the neural network architectures devised—to achieve a discernible decoded audio output from a low-density EEG headset, noting that there is one audio precprocessing method, and one EEG preprocessing method, which lead to objectively superior decoding than other approaches.

Keywords

EEG, brain-computer interfaces, imagined speech, neural decoding, stimulus reconstruction

[1] The dataset was compiled in collaboration with Jonathan Clayton [1]; with the exception of data collection, no other components were shared—all code, research, and writing is solely my own as presented in this dissertation, in support of satisfying the conditions of the MSc Speech and Language Processing with the University of Edinburgh.

Acknowledgements

With sincere thanks to my supervisors for their invaluable time and input; with especial thanks to my data-gathering comrade-in-arms who supported and put up with me when times got tough; with many thanks to the couple dozen participants who offered up their time and brain power for this project.

You've helped me complete my 5-year mission towards the MSc SLP.

Thank you.

Declaration

I declare that this dissertation was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Scott Wellington*)

Table of Contents

1. Introduction

Speech synthesis and recognition derived from neural signals alone: what was once constrained to technological forecasting is increasingly becoming a reality. With new machine learning algorithms and architectures, neuroscientific advances are now an explosively-active area of research for speech synthesis and recognition. This year alone has seen impressive results for the neural decoding of overt spoken and mimed utterances [2] [3], covert imagined utterances [4], and aurally-processed utterances [5]—it is these developments that have inspired the research presented here: particularly Akbari *et al.* [5], whose regression-based approach to decoding neural signals into vocoder features inspired the foundations of our investigation.

What is the purpose of such research and development, other than for its own sake? Augmentative and alternative communication devices (AACs) continue to be in development for individuals with speech and language deficits or impairments. Brain-computer interfaces (BCIs) are an exciting new direction to pursue for AACs; for instance, individuals with ALS (amyotrophic lateral sclerosis) or MND (motor neuron disease) frequently rely on eye-tracking technologies as a communication aid, however researchers are now exploring implanted BCIs to supplement or even replace these AACs [6] to improve words-per-minute communication. For individuals with locked-in syndrome, Neurolink's recent promises of high-fidelity, high-density BCIs for neuroprosthetic speech synthesis [7] can only be reified and peer-reviewed too quickly. In the meantime, research into speech synthesis from commercially-available and—crucially—non-invasive BCIs continues to gather steam; the research presented here focuses on the possibilities afforded from one such device: the Emotiv EPOC+ [8].

2. Background

2.1 Brain physiology and neuroimaging techniques

The brain is at once the most complicated and least understood organ of the body. Much that has become understood about brain mechanics results from observation of the *change* in mechanics following traumatic brain injury (TBI) [9], sequelae resulting from pre-natal development of brain lesions [10], or (inglorious) observations following experimental lobotomy or leucotomy procedures [11].

Fortunately various technologies now exist for functional neuroimaging of the brain, allowing researchers to not only view the utility of a brain structure from an aetiological standpoint, but to also view its processes *in real-time*.

2.1.1 EEG neuroimaging: advantages and disadvantages

Electroencephalography (EEG) is based on the observation that ”neurons wire together if they fire together” [12]; brain activity is therefore measured via electrical activity (for EEG) which results from ionic currents in the neurons when synapses ’fire’ (postsynaptic potentials). These affords high temporal resolution (to the order of $\sim$1 millisecond) but low spatial resolution (to the order of $\sim$5-9 centimetres) [13] [14], because it relies on measuring electrical potentials at the surface-level, and these potentials are prone to distortion from several interfering factors:

(i) Different tissues have differing densities which act as filters with differing conductivities, distorting the electric potentials conducted through them; the brain, meninges, cerebrospinal fluid, skull and scalp all lend to a mixed signal received by a surface-level electrode [14]. Notably for EEG measured non-invasively at the scalp, this signal is particularly distorted by the scalp itself [15].

(ii) Electrical potentials are not solely generated by brain activity; all myologic activity—tongue, facial muscles, heart, even environmental electrical activity—may create distorting artifacts within the signal. The most noise-inducing artifacts come from eye movement and blinking [16], which is involuntary, uncontrollable and occurs on average once every five seconds, with the resulting signal artifact being $\sim$100-400 milliseconds in duration [17]. While these artifacts can be removed with techniques such as independent component analysis (ICA) or wavelet transformation, these can also result in information loss [18].

(iii) EEG reference points are critical to measuring cortical electrical potentials, as an electrode measures the difference in potentials recorded between two positions: its own and its reference electrode [14]. Due to signal smearing and attenuation attributed to the above two factors, reference electrodes must be chosen that are neutral and not prone to distorting artifacts. Artifacts recorded by the reference electrode therefore contaminate the recorded signal at all other electrodes dependant on that reference [19]. Regrettably, reference electrodes are only selected for maximum neutrality, and are just as prone to confounders such as myologic activity. EEG signal smearing in this manner is especially exacerbated by participants with long and oily hair, which can conduct potentials across electrodes and reference locations [20].

Researchers—and individuals with speech impairments or deficits who would benefit from BCI devices—are constrained by cost, availability, and practicality. Foremost among these constraints is invasiveness: despite this year's promising brain signal-to-audio signal results from Akbari *et al.* [5] and Anumanchipalli *et al.*'s [3] electrocorticography (ECoG) research, such high-fidelity data is obtained through surgical procedure.

The advantages of non-invasive EEG-based BCIs are manifold: wearable EEG devices have the lowest overhead for both academician and consumer, and there is little set-up time required. In terms of data processing, deep brain sources contributing to speech and language processing (cf. §2.1.2) can be equally measured and utilised by scalp EEG for machine learning, as well as radial and tangential brain activity (within cortical gryri and sulci) otherwise unrecorded by other neuroimaging techniques [21]; further, improved data denoising and whitening methods have allowed low-cost, commercially-available EEG devices to perform on par with medical-grade devices (e.g. to train the widely-used P300 speller AAC from event-related potentials (ERPs) [22]).

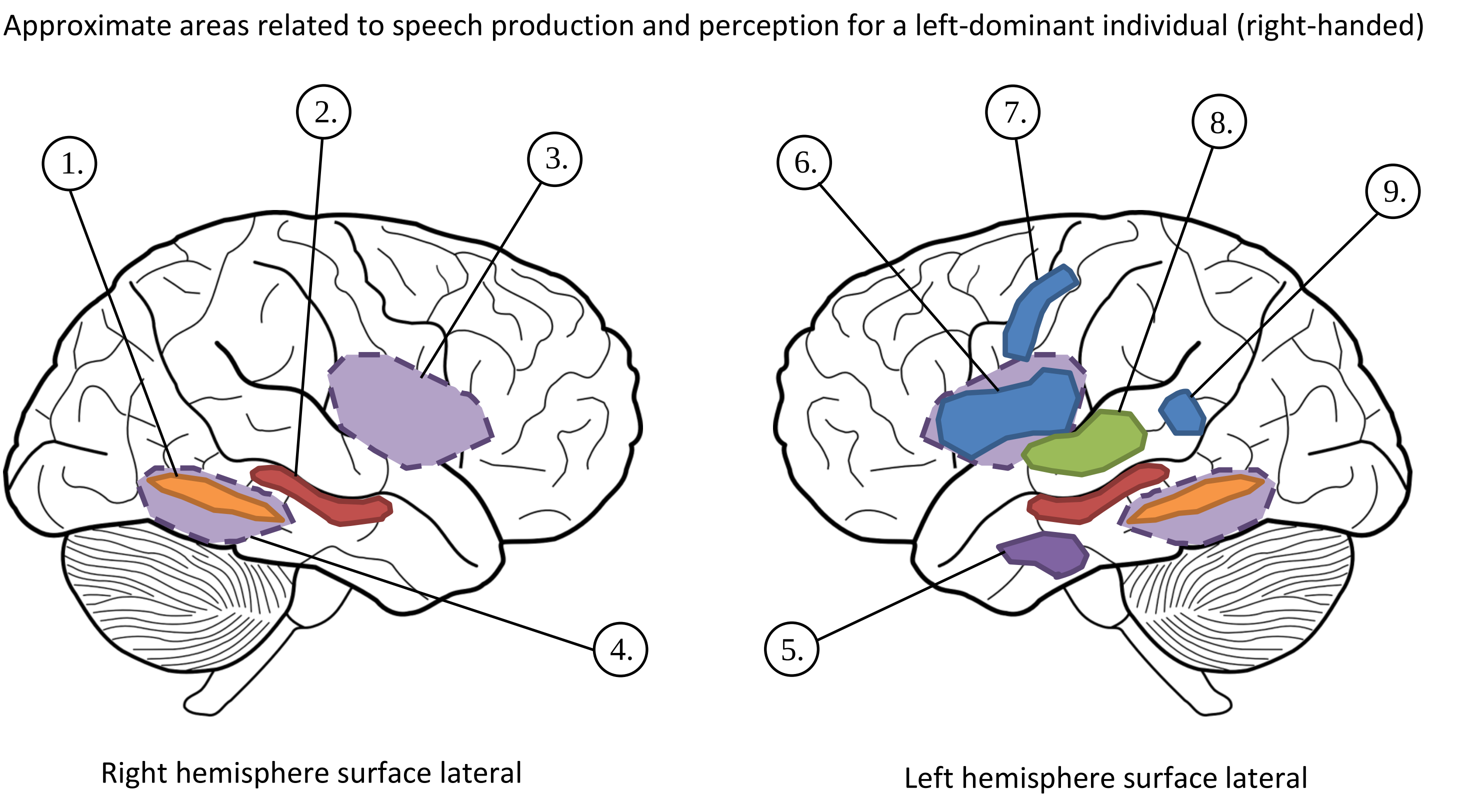
2.1.2 Speech and language processing in the brain

It is now well-established that the famous and eponymous Broca's and Wernicke's areas are only two of an interconnected and coordinated network of speech production and speech perception regions, illustrated in Figure 1.

A well-established observation is that right-handed participants overwhelmingly demonstrate language processing in the left hemisphere (∼96% have lateralisation as in figure 1), while left-handed individuals demonstrate a significant skew towards language processing in the right hemisphere (∼76% left-dominant lateralisation, ∼10% right-dominant lateralisation, and ∼ 14% bilateral [25]).

It is important to note that this observation, and those in Figure 1, are generalisations; each individual's brain physiology is different, which is an important consideration for machine learning models. It is tempting to believe therefore that user-dependent BCIs are equivocal to speaker-dependant automatic speech recognition (ASR) models.

However, while the 'holy grail' of low-cost EEG-based BCIs is for subject-independent devices that adapt their ERP or spectral model parameters to become subject-tailored (for instance, in the manner that speaker-adaptive ASR systems are currently under active development for dysarthric speakers [26] [27]), EEG signals demonstrate such a high degree of intra-subject variation within replica tasks that inter-subject variation is more robustly measured at an abstracted level (for instance, by the degree of sig-


Approximate areas related to speech production and perception for a left-dominant individual (right-handed)
1.
2.
3.
4.
5.
6.
7.
8.
9.
Right hemisphere surface lateral
Left hemisphere surface lateral

1. Superior temporal sulcus, involved in the interface between morphophonological and semantic processing
2. Superior temporal gyrus, involved in spectrotemporal analysis of auditory stimuli in speech perception
3. Insular cortex, involved in music processing, verbal memory, motor control and spoken emotion conveyance
4. Basal ganglia, involved in sequence processing (language planning) and sensorimotor control of articulators
5. Posterior and inferior temporal lobe, involved in memory and the lexical interface (sentential processing)
6. Broca's area, involved in sensorimotor learning, phonosyntactic comprehension and speech production
7. Precentral sulcus of the frontal lobe, involved in the audio-motor loop for articulation and speech production
8. Wernicke's area, involved in lexicosemantic processing, auditory word recognition and speech perception
9. Angular gyrus, involved in audio-visual processing (ventral) and auditory semantic processing (dorsal)

Figure 1: Approximated brain regions involved in speech and language processing, adapted from meta-analyses of neuroimaging research from Hickok & Poeppel [23] and Vigneau *et al.* [24]. Regions thought to be closely related in function for either speech production or speech perception are coloured the same. A dashed line indicates subsurface brain regions.

nal smearing or attenuation resulting from cranial tissue size, density and shape [14]; these have been recently shown to be reliable classification features for EEG biometric identification [28]).

2.2 Electroencephalography and low-density variants

2.2.1 Signal processing from EEG devices

Signals derived from EEG device recordings are high-dimensional, and the researcher must decide how best to process the (often extensive) data. If extracting features from the signal such as univariate statistics, values relating to spectral energy and entropy, etc., then for each participant, matrices will comprise [features extracted x time steps x device channels x frequency sub-bands].

Channel numbers for a typical research-grade EEG device range from 32 to 128, while intracranial ECoG devices typically comprise 256 channels [29]. Time steps depend on sampling frequency, which typically range from 256Hz to 512Hz, while ECoG devices typically have a sampling frequency of 1kHz. Finally, five EEG frequency bands are typically extracted for analysis: these are delta (<4Hz), theta (4–7Hz), alpha (8–15Hz), beta (16–31Hz) and gamma (>32Hz). These ranges are not formally codified, so a degree of overlap should be expected for these ranges in the literature [30]. Figure 2 illustrates the high-dimensionality of EEG data.

Unlike the task-specific variability of frequency sub-bands, the international 10-20 system is a standardised positioning (montage) of EEG surface (scalp) electrodes. Figure 3 shows the montage for the fourteen-channel Emotiv EPOC+ device [8].

2.2.2 The commercially-available Emotive EPOC+

The Emotive EPOC+ headset [8] used in our investigation was originally targeted towards gaming markets [30]. However, it has since gained traction as a research de-

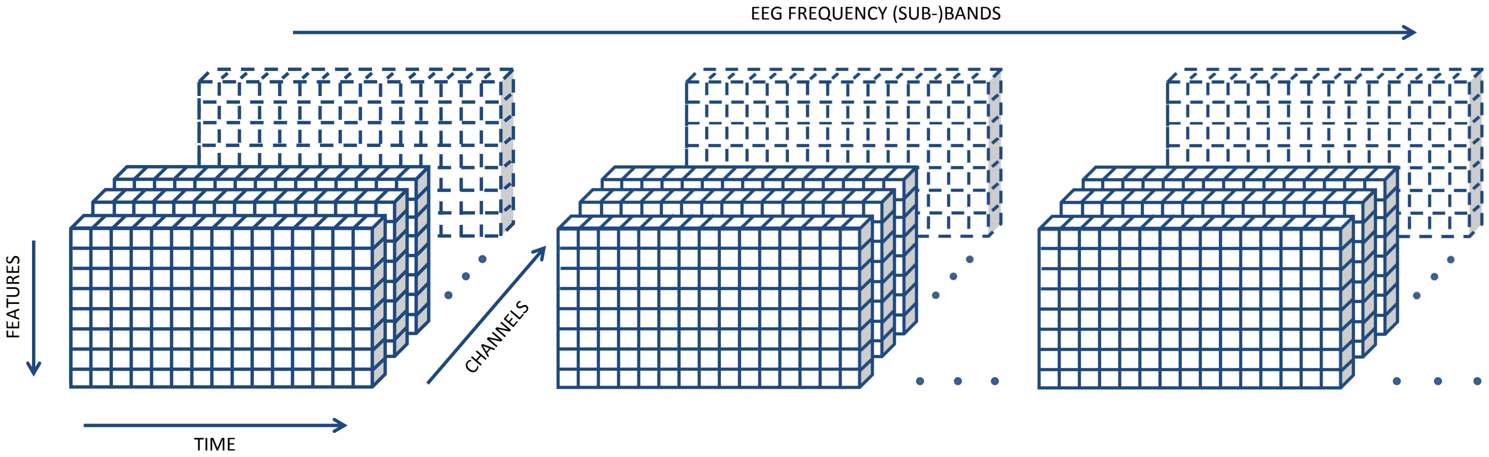


Figure 2: Illustration of the high dimensionality of EEG data, which is then subject to further signal processing.

vice to investigate the possibilities of achieving comparable results to higher-density, medical-grade EEG devices [31] [32].

The device uses saline-saturated felt sensors for scalp conductivity; for our investigation, we stream data at the device's maximum sampling rate of 256Hz. Notch filters at 50Hz and 60Hz guard the streamed signal against environmental electrical noise emissions (in the 60Hz range for, e.g. North America, and in the 50Hz range for, e.g. the United Kingdom) [20].

Figure 4 shows the Emotiv EPOC+ device, along with experimental considerations we discovered during the course of our data collection.

2.3 BCIs and machine learning for speech neuroprostheses: recent research and developments

Speech neuroprosthetic research has two divergent (but complimentary) areas: classification-based BCI applications (decoding continuously-valued EEG data into discrete, categorical values (words or phonemes) and regression-based BCI applications (decoding continuously-valued EEG data into continuously-valued output, e.g. audio).

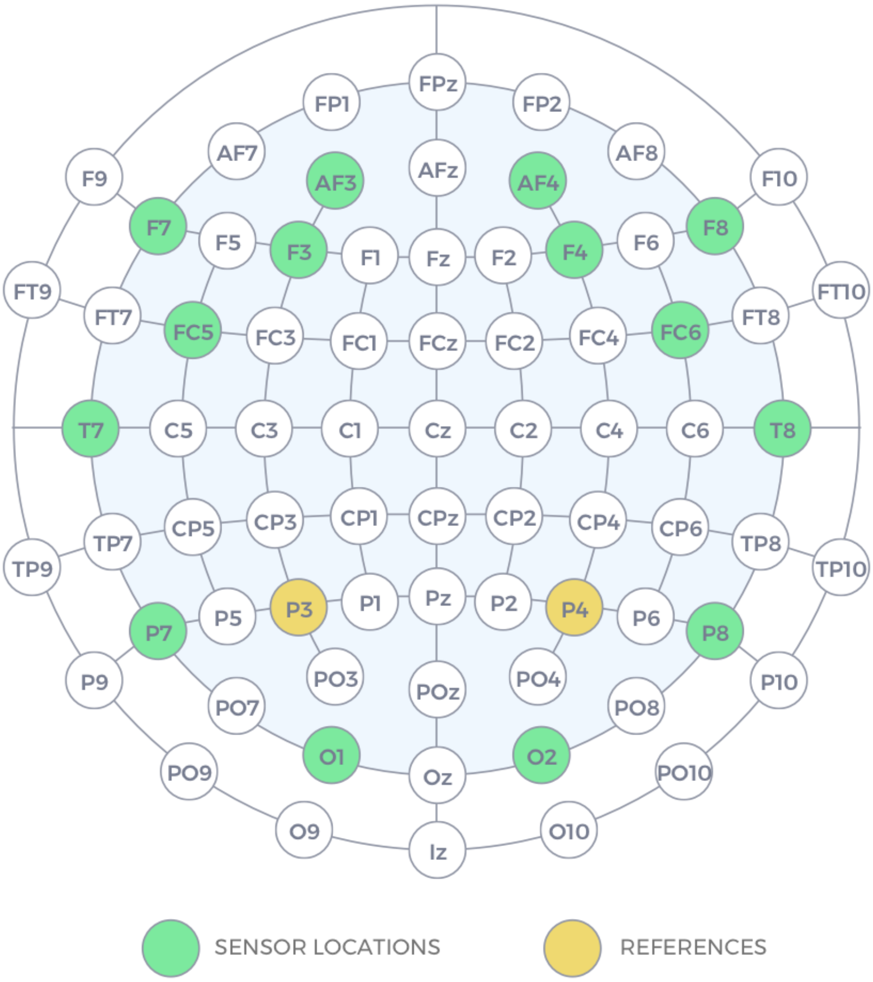


Figure 3: Electrode positions of the international 10-20 montage system: the top of the schematic is the front of the head (face). The fourteen electrodes and two reference positions of the Emotiv EPOC+ are highlighted (image from Emotiv [8]).

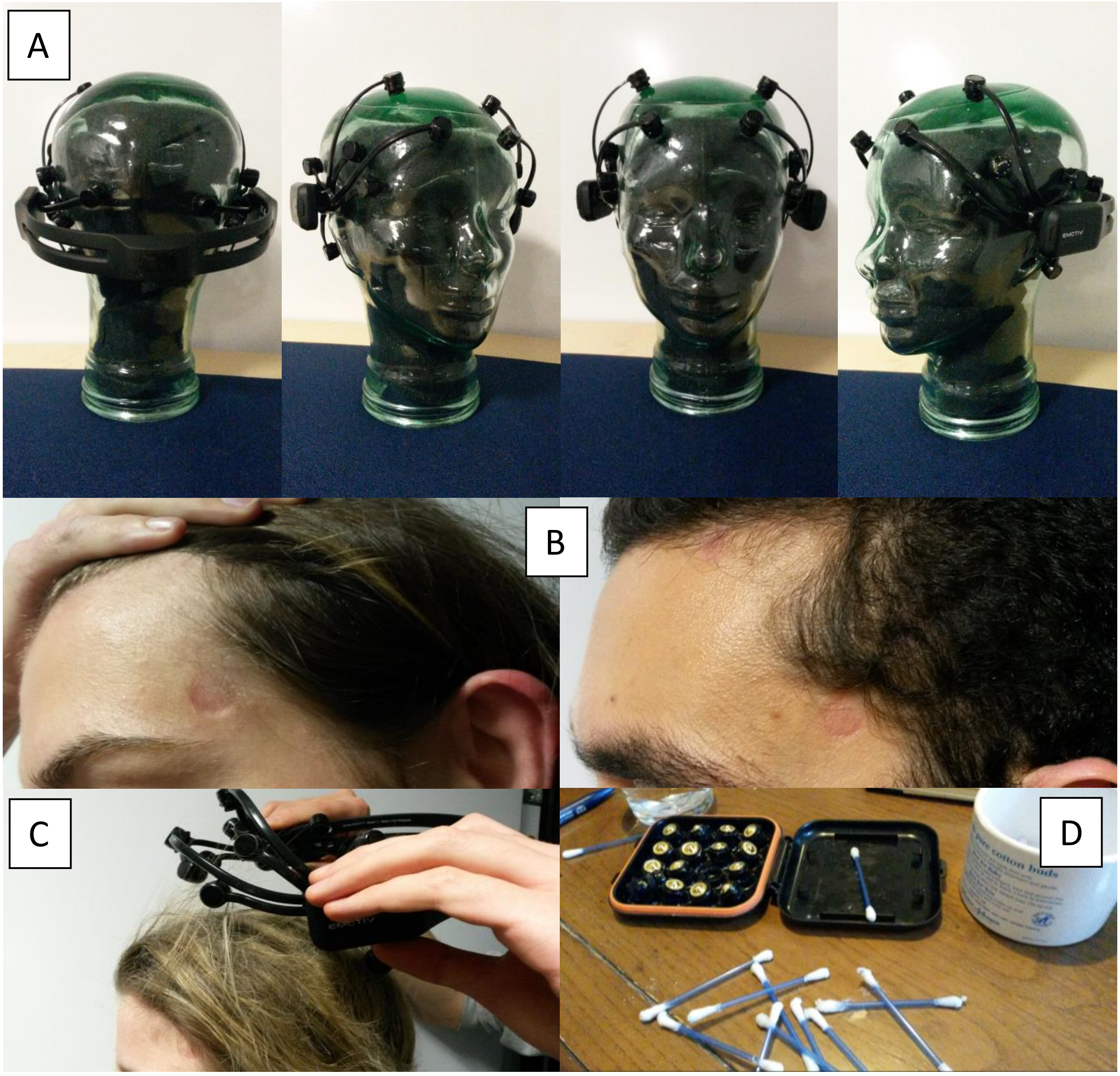


Figure 4: [A] The Emotiv EPOC+ [8] device. [B] Participants' scalps after using the device for 30-40 minutes; even after this short period, the pressure of the device can leave migraine-inducing recesses. [C] Long hair easily catches and winds around the sensors. [D] Electrodes require diligent cleaning and upkeep to prevent the saline-soaked felt pads from oxidising the gold-plated copper contacts.

Both approaches pose unique challenges; classification tasks typically window the EEG signal, and define a feature set whose values are then computed following transforms of the data from each window [30]. One of the benefits of this approach is that the high-dimensional EEG signal (cf. Figure 2) is transformed into sets of one- or two-dimensional arrays of a fixed size, which are easier for neural networks to extract and learn from higher-level features during training. Applications for these BCIs include the P300 speller, which allows the advanced user to communicate at $\sim$10 words per minute (with latest predictive text modelling) [3]; additionally, while promising developments have been made in whole-word classification (a binary selection within a limited number of classes), it has been noted that BCIs failing to reach above a consistent 70% accuracy threshold have limited real-world utility [33]: example EEG research this year for subject-independent binary word classification from imagined speech has variably hovered *under* this threshold: Cooney *et al.* [34] reports 62.37%. accuracy across multi-class words pairs, while Martin *et al.* [35] report 57.7% accuracy, noting that while classification accuracy can be as high as 100% for a single-class binary (e.g. 'yes' and 'no'), accuracy rapidly falls to chance levels (50%) as the number of word classes increases.

Regression tasks are also yet to reach 'real-world' utility levels, but this year has also seen impressive developments in decoding the EEG signal into audio output. These have have been researched in several directions:

2.3.1 Reconstructing speech from EEG of audio stimuli

Reconstructing speech from auditory stimuli requires decoding the brain activity underlying auditory processing and speech perception. Akbari *et al.*'s [5] recent research demonstrated impressive results using two decoding models; an autoencoder that encodes vocoder features into bottleneck representations before decoding out again, and an EEG decoder trained to decode into those bottleneck representations. Given the (relatively) streamlined nature of this decoding architecture, it is perhaps surprising that this was only proposed and investigated within the last year. Akbari *et al.*'s research and results form the inspiration for our investigation presented here: we investigate the extent to which is is possible to achieve comparable results on a (much) lower-density, non-invasive EEG device.

2.3.2 Reconstructing speech from EEG of imagined utterances

Reconstructing imagined speech requires decoding the brain activity underlying speech processing. It is arguably the most difficult EEG regression task (of the three presented here) because the neural processing of covert (imagined) speech is not sympathetic with the oral processing of overt speech in the time domain [36]; time taken to move articulators has no obvious neural correlative, resulting in a highly non-linear mapping between signals. For this reason, reconstructing speech from imagined utterances is typically constrained to a discrete-value decoding task, typically approached in the manner of automatic speech recognition problems [37] [38].

However, approaching the task from a discrete-value perspective has recently produced results potentially transferable to regression-based decoding; for instance, research into inner speech-related cortical activity has demonstrated that neural patterns correlating to specific phonetic stimuli (in natural speech) are still elicited when certain acoustic material is warped or degraded [36]. This 'phonetic masking' effect suggests separate processing for overt and covert speech stimuli, nevertheless operating in tandem. This is also supported by this year's research from Watanabi *et al.* [39] who demonstrate that the spectral envelopes for overt and covert speech are temporally correlated, reporting that neural oscillations and speech periodicity are synchronised phenomena.

2.3.3 Reconstructing speech from EEG of spoken and/or mimed utterances

Reconstructing speech exclusively from EEG signal is uniquely highly contaminated with electrical potentials from articulator movement and voicing. While Zhao & Rudzicz [40] present this as a confounding factor, this year's research from Angrick *et al.* [2] has shown that the resultant 'mixed' signal can fortify a neural network's ability to learn higher-level features from the data, with impressive audio reconstruction of spoken utterances using deep convolutional neural networks (CNNs) and the raw EEG signal. Likewise, Anumanchipalli *et al.* [3] demonstrated impressive audio reconstruction of mimed utterances by using acoustic-articulatory inversion mapping, and using this kinematic representation as an intermediate stage within the neural decoder model to predict acoustics. This extra layer of representation within the decoder model served in part as inspiration for our stacked decoder model presented here (cf. §4.2).

3. Motivation

The number of publicly-available EEG data for imagined speech is vanishingly small. Many researchers have collected such datasets (e.g. [41], [42]), but not released it publicly.

We were only able to locate two publicly-available EEG datasets for imagined speech. These are:

- The KARA ONE dataset from Zhao & Rudzicz [40] (imagined phonemes/syllables and short words from 14 participants).

- The unnamed dataset from Nguyen *et al.* [43] (imagined phonemes/syllables and short words from 15 participants).

As Nguyen *et al.* note, the scarcity of publicly-available datasets is a severe dampener to researchers' ability to reproduce results, and explore alternative approaches and algorithms.

Therefore, for this investigation, we compile a new EEG dataset for imagined speech (along with heard speech and overt speech conditions), comprising 22 participants' worth of data (cf. §4.1). This we call the 'Fourteen-channel EEG with Imagined Speech' (FEIS) dataset [44], publicly released under the ODC-By licence. To our knowledge, this is the first such dataset compiled for imagined speech using a low-density EEG headset. It is our hope that FEIS will allow future researchers to explore comparative models, and improve upon the findings presented in our investigation.

3.1 Primary Goals

Having compiled the FEIS dataset, we may ask the question: what can we do with it?

As established in §2.3, many researchers have promisingly reconstructing speech from EEG data. However, these results have all derived from high-density EEG devices (often medical-grade intracranial ECoG devices); very limited research exists replicating such experimental findings using a low-density EEG device—specifically, a commercially-available device like the Emotive EPOC+ [8] (cf. §2.2.2).

Before proceeding, our (first) preliminary investigations were intended to establish some 'baseline' expectations for model performance, and gain some intuitions about the results expected; we therefore explored classification models as applied to The KARA ONE dataset [40].

3.2 Preliminary results from classification models and the KARA dataset

Using a deep convolutional neural network (CNN) model architecture (Deep4Net) within the BrainDecode deep learning EEG toolbox [45], we explored a binary 'vowel vs. consonant' classification task. Figure 5 presents the plotted results of this preliminary investigation. As can be seen from the probability graphs and scalp plots, the model struggles to classify either category better above 50% chance.

Recent studies from Schirrmeister *et al.* [45] and this year's research by Angrick *et al.* [2] demonstrated that deep CNNs are able to train from the raw EEG signal (without preprocessing the data), as higher-level features are able to be learned from lower-level features extracted in the network's initial layers. This has the advantages of being less computationally expensive, having less parameters to optimise, and avoiding informa-

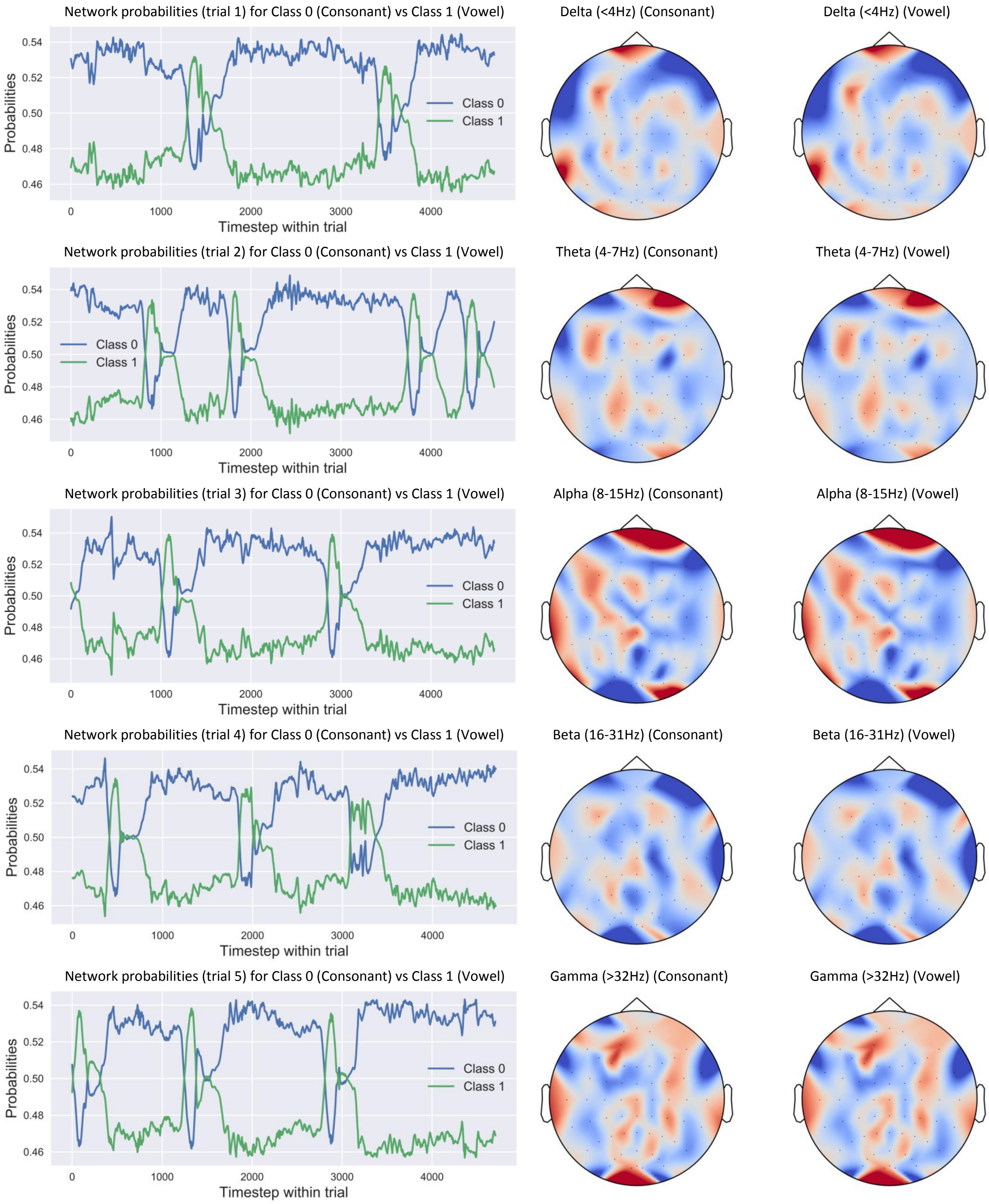


Figure 5: Classification results plotted from MNE [46] [47] using Deep4Net [45] for 'vowel vs. consonant' classification from imagined speech EEG data. In all trials (left), the model was unable to delineate the two classes, nor determine distinguishing features between the five EEG frequency bands (right).

tion loss through data preprocessing (as can result from ICA if pollution from signal artifacts leads to rejection of ’contaminated’ signal that otherwise contains data the network could still learn from [48]).

While our preliminary results do not support these assertions, Clayton [1] reports promising results for CNN vowel/consonant binary classification using the FEIS dataset [44].

3.3 Hypotheses for regression models and the FEIS dataset

Provided the preliminary results and considerations in §3.2, we establish three hypotheses:

(1) The limited number of channels on the Emotiv EPOC+ headset [8] supply data that is possible to decode as a restrained-category classification task, but is too low density to suitably decode as a regression task.

(2) Where objective measurements are possible (and meaningful) for decoded output, the ’hearing’ condition in our trials (cf. §4.1.1) should outperform the ’thinking’ condition (as supported by previous research; cf. §2.3.2) and the ’speaking’ condition (following Zhao & Rudzicz’s [40] discovery that overt speech severely introduces signal artifact distortion).

(3) Where decoded audio output is obtained from a regression task using FEIS data, it will not be achieved through modelling our neural network architectures on research from Akbari *et al.* [5] alone (cf. §2.3.1); the denoising stacked decoder model we employ (cf §4.2) should achieve better objective metrics.

3.4 Preliminary results from regression models and the FEIS dataset

Given our experiment parameters and the number of model permutations under investigation, we use the FEIS [44] data of the participant with the highest average validation accuracy as reported by Clayton [1] for all experimental procedures. In this regard, we are exclusively testing our hypotheses under a speaker-dependant (or 'thinker-dependant') framework.

Without preprocessing of EEG or audio data, EEG signal for the participant (participant '19' in the FEIS dataset [44]) was firstly decoded for the three conditions under investigation, using a fully-connected neural network decoder as outlined in §4.2. These results from the raw data formed our baseline measures for comparison, and are presented in Table 1.

	Hearing	Speaking	Thinking
/i:/	3.68,-1.19	3.92,-1.79	4.30,-3.14
/æ/	4.50,-4.17	4.13,-2.41	4.46,-3.63
/u:/	3.97,-0.95	4.23,-1.50	4.59,-2.65
/ɔ:/	4.18,-0.76	4.39,-1.12	4.74,-2.09
/f/	**3.08**,-5.96	**3.25**,-6.90	**3.59**,-8.83
/v/	3.56,-2.10	3.78,-2.88	4.17,-4.60
/s/	3.36,-5.91	3.32,-5.67	3.66,-7.59
/z/	3.91,-3.97	3.68,-2.83	4.01,-4.27
/ʃ/	3.15,-3.41	3.29,-4.06	3.64,-5.94
/ʒ/	4.04,-2.46	4.19,-2.92	4.56,-4.57
/p/	4.23,-2.30	4.51,-3.41	4.87,-5.04
/t/	4.03,-1.56	4.14,-1.78	4.49,-3.07
/k/	4.00,-1.99	4.14,-2.36	4.50,-3.86
/m/	4.06,-4.71	4.30,**-0.95**	4.68,**-1.93**
/n/	4.04,-1.16	4.23,-1.58	4.58,-2.74
/ŋ/	4.12,**-0.73**	4.34,-1.09	4.73,-2.18

Table 1: Preliminary results from regression models (cf. §4.2) using non-preprocessed (raw) EEG signal and WORLD [49] audio features, for the 16 phonemes under investigation. As compared to the original audio stimuli, Log spectral distortion (LSD) results are provided in red; signal-to-noise ratio (SNR) results are provided in blue. Best results for each metric are in bold.

4. Methods

4.1 Data collection

EEG data for our FEIS dataset [44] were collected from 22 participants; two of these participants' data were collected for investigation into Chinese tone decoding (c.f. Clayton [1]), while twenty-one of the participants' data were collected for investigation into English phoneme decoding (one of the participants provided both Chinese and English EEG data in separate trials).

Participants for BCI research tend to be excluded if they are left-handed [25]; as discussed in §2.1, this is due to the skew towards heterogeneous speech and language processing regions of the brain when modelling cortical activity between left- and right-handed individuals. However, this does result in a dearth of publicly-accessible EEG data for left-handed individuals. Our intention was to accrue as much data as possible to fill in the database 'gap'; therefore we had no exclusionary criteria for participants. Of our twenty-one English-speaking participants, two are left-handed, one is ambidextrous, and ten are non-native speakers of English (FEIS metadata is available at [44]).

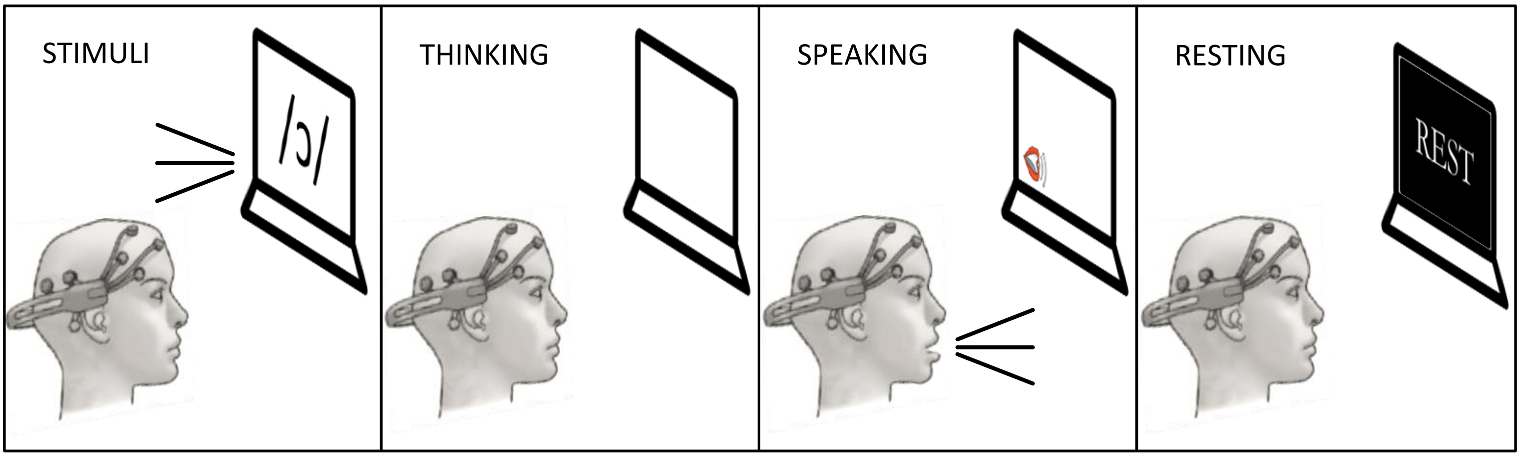


Figure 6: Sequential illustration of the 5-second EEG-recording stages for each trial, for each participant. In a hemi-anechoic environment, participants repeated 160 iterations of this sequence (10 repetitions of the 16 phonemes under investigation). Image of individual wearing the Emotive EPOC+ [8] sourced and adapted from Abhang *et al.* [30].

4.1.1 Trial design and prompt selection

The sixteen English phonemes under investigation were chosen to represent the largest categorical spread of contrastive phonological features (e.g. [±nasal], [±back], [±voice], etc.), while not being exhaustive. These are:

/i/ /u:/ /æ/ /ɔ:/ /m/ /n/ /ŋ/ /f/ /s/ /ʃ/ /v/ /z/ /ʒ/ /p /t/ /k/

Audio recordings of these phonemes were taken from each participant, as spoken in isolation. It should be noted that for the phonemes /p/, /t/ and /k/ participants were instructed to form a neutral release (resulting in /pə/, /tə/ and /kə/). Although it is valid to measure and decode EEG for unreleased plosives (with impressive results from Zhao & Rudzicz [40]), for the purposes of a regression-based decoding task, we required an audible phoneme such that objective metrics used with the decoded audio output would provide meaningful results.

Figure 6 illustrates the trial procedure and its four stages. Once the participant begins the trial, an OpenViBE [50] scenario automatically plays all audio-visual stimuli through to completion. In this regard, the experiment is fully automated; this is necessary because participants conducted the trial in a hemi-anechoic environment and in isolation to minimize EEG signal artifacts resulting from audio or visual activity external to the experiment. Participants were also instructed to remain still and relaxed during the first three stages ('stimuli', 'thinking' and 'speaking'). Our aim in this regard was to control for EEG signal artifacts, such that data processing algorithms would be more successful in isolating uncontrollable EOG (electrooculogram) signal noise from blinking. In the 'resting' stage, participants were instructed to relax and de-focus; while this also relieves cognitive load from participants, it also provides a 'resting state' measurement which can be used for task-specific feature extraction and learning EEG markers [15] [16].

The experiment design was adapted from Zhao & Rudzicz [40], with the following key differences:

- Between our stages 'stimuli' and 'thinking', Zhao & Rudzicz [40] have a 5-second 'prepare articulators' stage. We believed that this would result in EEG signals polluted with strong articulator movement artifacts; we therefore replace this stage with a one-second 'fixation point' (following the cross-and-bullseye design recommended by Thaler *et al.* [51]) to fix participants' gaze toward the OpenViBE [50] scenario and prevent signal artifacts resulting from eye-movement saccades and fixations.

- Zhao & Rudzicz [40] present whole-word visual prompts to their participants. However, in addition to minimizing the sources of noise artifacts, we also wished to control for data-polluting ERP responses resulting from any experimental design component. To this end, we chose non-word audio-visual prompts because word recognition typically prompts an ERP response, attributed to the accociationist account of phonology processing [52].

- Additionally, as opposed to having the same acoustic stimuli played to all participants, we chose to have the phoneme stimuli recorded in the participant's own voice and digitally repeated five times in succession (as opposed to the participant performing five renditions) because ERP responses can also arise from unusual or 'thought-provoking' phonetic stimuli [53].

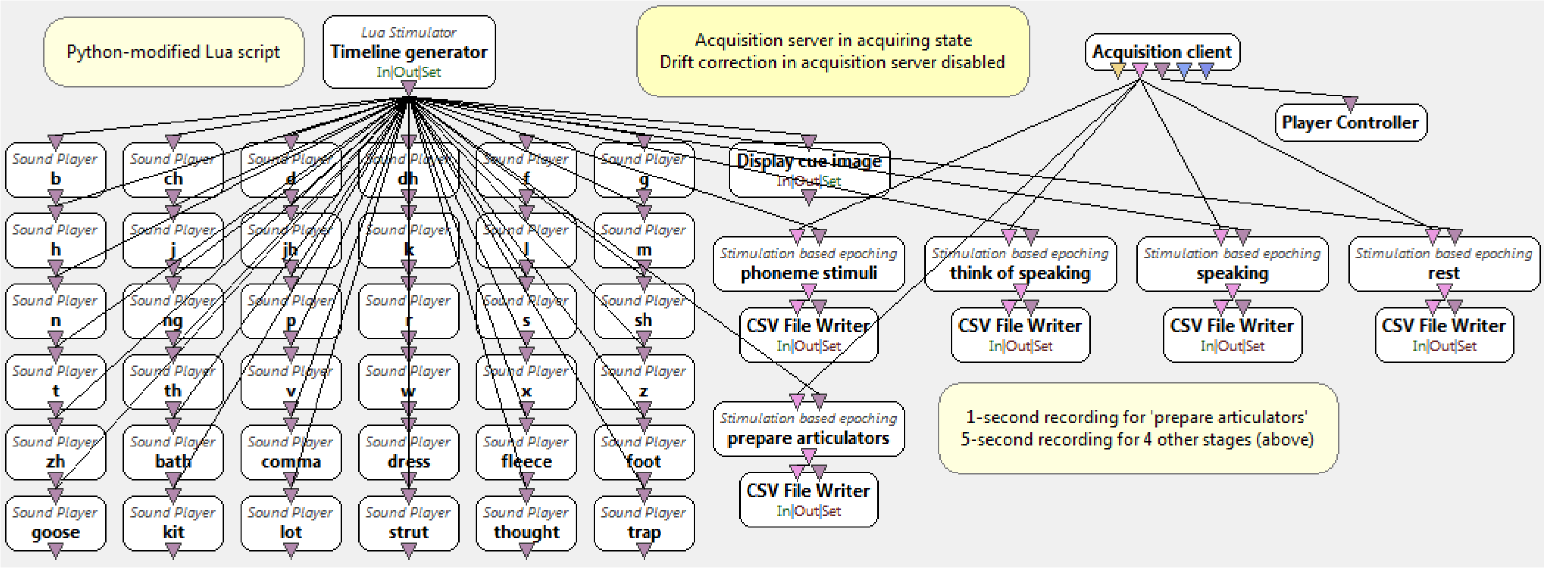


Figure 7: The OpenViBE [50] scenario used in our trials, configured to conduct the experiment with a randomised, user-configurable combination of thirty-six English phonemes.

4.1.2 Audio recording

Original audio stimuli was recorded from participants at 44.1kHz using a DPA cardiod (4088) microphone connected via an XLR-to-USB signal adapter. During each trial, the participants heard their own voice repeating one of sixteen recorded phonemes; each participant's phoneme was repeated five times in precisely five seconds. A script was written to record all sixteen spoken phonemes in one short sitting; using SoX [54], this automatically processed a participant's single phoneme recordings into a single second window, duplicated 5 times, and downsampled to 16kHz.

4.1.3 EEG recording

EEG was streamed from the Emotiv EPOC+ device [8] at 256kHz via TCP stream using the CyKIT Python 3x server [55] to the OpenViBE [50] acquisition server.

Figure 7 shows the Lua-scripted OpenViBE scenario we created to play the audio-visual stimuli and record EEG data. Lua scripts epoched the data for each 5-second stage of our trial. For each of the sixteen phoneme under investigation, the four-stage sequence was repeated ten times. This was conducted in ten randomised blocks of sixteen phonemes. The resulting 160 epochs took ∼58.5 minutes to record; we therefore split the trial over three sessions to relieve participants' cognitive fatigue.

Sensor positioning was standardised for all participants via positioning the reference electrodes at P3 and P4 to rest on the mastoid process behind each ear; sensor arms were visually inspected and adjusted to be symmetrically-placed. Before recording, 100% conductivity for all electrodes was confirmed via the EmotivBCI application [8].

4.2 Experimental design and model architectures

As noted by Takaki & Yamagishine [56], one of the advantages of vocoder features (as opposed to Mel-frequency cepstrum coefficients) is that the spectral envelope extracted tends to be very accurate and stable; such reliability makes vocoders such as STRAIGHT [57] and WORLD [49] good choices for deep neural network (DNN) speech synthesis models to learn from.

Akbari *et al.* [5] employ WORLD vocoder features for their research into decoding heard stimuli from ECoG data; it is this research that inspired our machine learning model design, with some adaptations. It should be noted that Akbari *et al.* [5] have not provided their code as a public release (it is available on request); therefore, we designed and created the neural network architectures as bespoke for this investigation, using PyTorch [58]. There are several network parameters which were inferred from Akbari *et al.* [5] as they were not made explicit, which we will highlight.

Figure 8 is a diagram illustrating our employed neural network architectures. Three different models were scripted: an EEG signal decoder; a WORLD feature autoencoder; and a stand-alone decoder. The stand-alone decoder (the 'stacked' decoder) optionally decodes the auto-encoder output into a denoised version of the same (normalised) WORLD features), and is novel to our investigation..

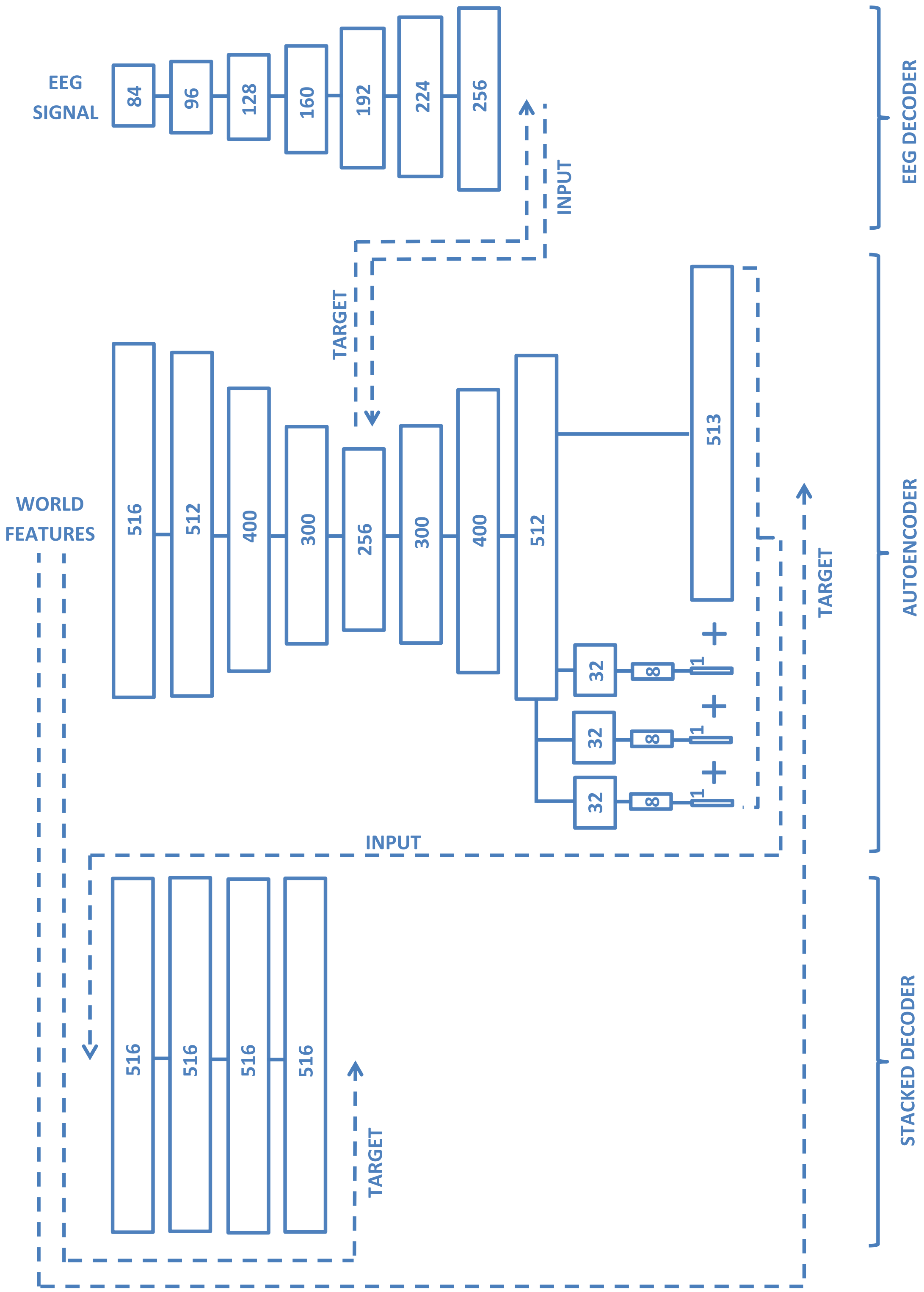


Figure 8: DNN neural network architectures used in our investigation: an EEG signal decoder; a WORLD feature autoencoder; and a stacked decoder model. The length of the arrays processed by each layer is provided, along with data input/target labels expressing the relationship between each model in our EEG-to-WORLD features pipeline.

Akbari *et al.* [5] investigated using a fully-connected network (FCN) model to decode the EEG signal, alongside a model which concatenates the FCN output with the output from a locally-connected network (LCN). They report ESTOI measures of .45 and .47 respectively when the models are trained on WORLD vocoder features. When trained on spectrogram features, the reported ESTOI measures are .40 and .40 respectively. As the results are comparable, we decided to pursue an EEG decoder model with the lowest complexity, while still allowing for comparable results; because we compare twenty-seven permutations of this model (see below), we decided this would be practical in terms of computational cost.

The autoencoder employed by Akbari *et al.* [5] was trained on 80 hours of speech from the Wall Street Journal CSR corpus [59]; as this corpus is licensed and not freely available, we instead trained our autoencoder on CSTR's VCTK corpus [60] from the University of Edinburgh. The VCTK corpus comprises 44,242 audio files, with $\sim$44 hours of English speech from 109 speakers. Adapting Python code from Watts [61], we are able to interpolate f0 values from WORLD feature arrays extracted from PyWorld-Vocoder [62]. These then provide four arrays of WORLD feature values per sample of audio: the spectral envelope, (mean) band aperiodicity, f0, and a voiced/unvoiced 'excitation' label. Concatenating these values provides a 516-value array of WORLD features which our autoencoder is trained on, for each sample, for each audio file. As in Figure 8, the autoencoder output branches into these four components, each of which has their own graph of learnable weights updated independently when backpropagating the loss (mean squared error) through the network.

The encoded 256-value bottleneck features from the autoencoder are used as the target output to train the EEG decoder model. These bottleneck features are subject to additive Gaussian noise to make the autoencoder more robust to signal artifacts; Akbari *et al.* [5] do not specify the parameters of their additive Gaussian noise, so we followed the recommendations of Jin *et al.* [63] to incorporate this within our model.

Between all network layers we observe the activation functions of Akbari *et al.* [5]: LeakyReLU for all hidden layers, ReLU for the 513+1+1+1 output arrays of the autoencoder, and tanh for bottleneck feature and EEG decoder outputs. We also adopt tanh as the output for our stacked decoder model. Preliminary results for our EEG decoder indicated that network performance suffered when using dropout; Akbari *et* *al.* also note that this was the case with their autoencoder, choosing to replace the functionality of dropout with the additive Gaussian noise; subsequently, none of our final models have dropout (Akbari *et al.* retain it for their EEG decoder). All models were trained using the Adam optimiser [64] on the Eddie Mark 3 compute cluster [65] at the University of Edinburgh.

All our models were trained using an 80/10/10 training/test/validation split. As regards the EEG data, due to limited data (10 recorded epochs for 16 phonemes) we took one epoch (at random) from each phoneme for our held-out set; however, the full set of ten epochs was still used to form an averaged, denoised EEG signal for each phoneme under each condition (hearing, speaking and thinking), then used to train the models.

We investigate three audio data preprocessing methods and three EEG data preprocessing methods (below). For our full investigation, therefore, we trained three autoencoders (one per audio preprocessing method), nine stacked decoder models (hearing, speaking and thinking × three audio preprocessing methods), and twenty-seven EEG decoding models (hearing, speaking and thinking × three audio preprocessing methods × three EEG preprocessing methods).

4.3 Data preprocessing

Although DNNs are universal function approximators, finding the correct data preprocessing method may significantly impact a network's ability to learn from the data. As noted by Choi *et al.* [66], a network may be able to represent a complex function, but this does not arguably equate to a network's ability to learn a complex function. Despite the same signal transforms applicable to either EEG or audio data (e.g. fast Fourier transform (FFT); discrete wavelet transform (DWT); periodicity transform (PT), etc. [67]) the preprocessing methods typically applied to audio and EEG data are particular to those domains, and the chosen method applied is usually also task-specific within that domain [66] [68].

Data scaling via normalisation or standardisation for real-valued input and output features is particularly important for feature sets that comprise data with different units of measurement (as with our WORLD features); neural networks usually train best when features are scaled to be within the same comparable range. This is also true of audio and EEG spectra whose values have a non-normal distribution; thus, we want to preserve the underlying distribution of the data while ignoring the unit variance [68]. The discussion surrounding the best preprocessing technique for differently-valued data is ongoing[66].

4.3.1 Audio data

The 44,242 audio files of the VCTK corpus [60] were downsampled to 16KHz (to match the sampling rate of our participants' audio) using the Sound eXchange (SoX) audio editing utility [54], configured to dynamically adjust gain to guard against clipping samples during resampling, and removing all silence from the beginning, middle and end of the audio (defined as 1% of the maximum value of the sample value) beyond 0.1 second. This greatly compresses the size of the corpus audio files from ∼14.1 GB to ∼2.28 GB. 8 files were determined by these SoX parameter threshold values to be comprised solely of silence, and were discarded.

Once all corpus .wav files were converted into .csv WORLD feature files, scripts were created to perform outlier removal and three normalisation methods. Outlier removal was performed using the '68-95-99.7' rule: WORLD feature values outwith three standard deviations from the norm were discounted from training. Finally, features were normalised into separate files using three methods:

(a) Min-Max normalisation. It has become common practice to scale values to within the range [0.1, 0.9] for neural networks [69], and this is the approach taken in our investigation (4.1), where a is the minimum scale (0.1) and b is the maximum (0.9).

$$x_i{}' = \frac{(x_i - \min(x))(b-a)}{\max(x) - \min(x)} + a \tag{4.1}$$

Min-Max normalisation is also employed by Akbari *et al.* [5]. However, preliminary results showed that our models only slightly outperformed results from the raw data (Table 1). Min-Max normalisation may not always be the best choice, as it is also sensitive to outliers.

(b) Z-Score standardisation (also ambiguously known as z-score normalisation). Here the WORLD features are rescaled to approximate a standard normal distribution.

$$x_i{}' = \frac{x_i - \bar{x}}{\sigma} \tag{4.2}$$

Z-Score standardisation is an integral data processing technique within machine learning [70]. However, as the data is re-centred around zero, it is possible that the neural network outputs may be sparser (with more zero-valued fields predicted). Z-Score standardisation is the method used for Watanabi *et al.*'s research into imagined speech (cf. §2.3.2).

(c) Box-Cox transformation. Similar to the Z-Score, it aims to expand low values and compress high ones, making the distribution more Gaussian. The transformed values are scaled to the desired interval, dependant on how the zero values and the maximum values are transformed during Gaussianisation [71]. The lambda parameters for this transform are estimated from the WORLD feature data, and were automatically calculated using the SciPy package [72].

$$x_i{}^{(\lambda)} = \begin{cases} \frac{(x_i+\lambda_2)^{\lambda_1}-1}{\lambda_1} & \text{if } \lambda_1 \neq 0 \\ \ln(x_i+\lambda_2) & \text{if } \lambda_1 = 0 \end{cases} \tag{4.3}$$

As with Min-Max normalisation and Z-Score standardisation, the Box-Cox transformation can also be applied to EEG signal data [73].

4.3.2 EEG data

Following EEG signal preprocessing, Akbari *et al.* resample their EEG data to 100Hz. The WORLD vocoder [49] used in our investigation is configured to create feature sets of 516 values at a sampling rate of 200 per second for 16KHz audio; therefore, to approximate a sample-for-sample mapping for regression, the 256Hz EEG signal was resampled to 200Hz using OpenViBE's [50] low-pass FIR filter to prevent spectral aliasing. Data were then triplicated to be processed using three approaches:

(a) Band frequency filtering & Hilbert transform

As in §2.2.1, EEG signal data is typically bandpass-filtered to isolate frequencies associated with speech and language processing. Typically, only selected ranges will be used for machine learning models; for instance, neural activity has been demonstrated to be time-aligned to the duration and pacing of syllables in theta frequency bands (4-8Hz) and of whole words or short phrases in the delta band (1-4Hz) [74]. DNNs that have been trained to investigate speech rhythm neural processing at these filtered frequency bands use the spectral envelopes for feature extraction [75], which can be extracted using the Hilbert transform.

Akbari *et al.* [5] extract the spectral envelopes for brain frequency components below 50Hz, but do not specify what these are; these are then averaged with eight high gamma envelopes between 70-150Hz to create a 256-value array (one value per channel) for each time step, which their EEG decoder is trained on. We approximate this process by extracting the envelopes from the five EEG frequency bands (delta, theta, alpha, beta and gamma) and concatenating these with the averaged eight high gamma envelopes between 70-150Hz. This creates an 84-value array (six values per channel) which we use to train our EEG decoder. We performed bandpass filtering and the Hilbert transform using the OpenViBE [50] environment.

(b) Independent Component Analysis (ICA)

ICA is signal decomposition to find the independent componants comprising it; this can then be used to isolate and remove signal artifacts resulting from electrooculogram (EOG) or electrocardiogram (ECG) noise. The benefit of using ICA over other regression-based removal of EOG or ECG artifacts is that ICA does not require a reference signal. However, ICA does risk removing brain data alongside EOG/ECG data, and this is a problem which is exacerbated with the fewer number of channels available [76].

Lau *et al.* [77] note that while the minimum number of channels necessary for data-preserving ICA can vary depending on the task, at least twenty channels are needed. The Emotiv EPOC+ [8] used in this investigation has fourteen channels; Lau *et al.*'s observation appears to be confirmed by Clayton [1], who found that using ICA with Emotiv EPOC+ data worsened baseline performance of classification models.

We performed ICA with MNE [46] using the MNE-Python package [47] for artifact correction. As ICA is (very) computationally expensive, we use FastICA (a form of simplified whitening) for comparable results [76]. After ICA is computed on our EEG data for EOG removal, the data is then processed using band frequency filtering & the Hilbert transform, as above.

(c) Surface Laplacian. A scalp surface Laplacian transform is a common technique in EEG data processing to improve spatial resolution, and lessen the impact of distortions resulting from signal artifacts near reference electrodes polluting the entire signal [78]. However, it is usually performed on high-density EEG devices, as the transform 'smooths' the signal from several nearest-neighbour electrodes into a single-signal transformed output [79].

We performed a surface Laplacian transform using the OpenViBE [50] environment, transforming the left-hemisphere nearest-neighbour electrodes at F3, FC5, AF3, F7 and T7, and the right-hemisphere nearest-neighbour at T8, F8, AF4, FC6 and F4 into two transformed signals which were then processed using Band frequency filtering & the Hilbert transform, as above.

5. Results

For the nine permutations of models investigating the 'hearing' condition (Table 2), /m/ was discovered to have the lowest LSD of all repeated-phoneme stimuli under investigation (when compared to the original audio stimulus; LSD=1.61, SNR=-0.83); for the 'speaking' condition (Table 3), /u:/ was discovered to have the lowest LSD (LSD=1.57, SNR=-0.93), and /u:/ was also discovered to have the lowest LSD under the 'thinking' condition (LSD=1.69, SNR=-0.89) (Table 4). In all instances, the combination which achieved this lowest LSD was MinMax preprocessed audio and bandpass-filtered EEG with the average envelope from the Hilbert transform.

When using the stacked decoder model (Table 5), /f/ is the minimum for LSD in the hearing, speaking and thinking conditions (with Box-Cox (LSD=2.73, SNR=-6.58), Box-Cox (LSD=2.70, SNR=-6.54) and Min-Max (LSD=3.00, SNR=-6.25) audio preprocessing, respectively). However, /f/ is also the minimum for LSD in the hearing, speaking and thinking conditions with the non-preprocessed raw data (cf. §3.5). This is likely because the audio is ostensibly white noise; indicatively, the respective SNR scores for each measure are the highest in their field for each condition, highlighting the importance of having different objective metrics (here LSD and SNR) which quantify different transforms of audio signal [80], due to very high noise distortion involved with research into brain decoding via regression.

Universally, the SNR is a negative value; with the power (or energy) of the waveform measured in dB, we interpret this as being that the noise in the decoded signal has greater energy than the signal itself. For the nine permutations of models investigating each of the hearing, speaking and thinking conditions, /v/ was found to have the lowest SNR of all repeated-phoneme stimuli under investigation (-0.38 with Min-Max and Laplacian; -0.36 with Min-Max and Laplacian; and -0.36 with Min-Max and ICA, respectively). When using the stacked decoder model, /m/ has the lowest SNR for hearing and speaking (-0.89 and -0.83 with Box-Cox preprocessing), and /ŋ/ for thinking (-1.45 with Z-Score preprocessing).

When compared to the raw (non-preprocessed) data decoding (Table 1), we can see that Min-Max audio processing outperforms raw data for every phoneme stimulus, for all conditions (Tables 2, 3, 4). Box-Cox and Z-Score audio preprocessing performed notably worse in all conditions. The best EEG preprocessing technique, in all conditions, remains the bandpass-filtered EEG with the average envelope from the Hilbert transform (mean hearing LSD=2.17, SNR=-1.61; mean speaking LSD=2.14, SNR=-1.59; mean thinking LSD=2.18, SNR=-1.60). The best results recorded here are from the audio and EEG preprocessing techniques employed by Akbari *et al.* [5] (c.f. §2.3).

Figure 9 visually illustrates the differences the between original, decoded (without stacked decoder) and decoded (with stacked decoder) audio stimulus which achieved the lowest LSD under the 'hearing' condition (/m/). Figure 10 illustrates likewise for the 'speaking' condition (/u:/) and Figure 11 for the 'thinking' condition (also /u:/). Figure 12 illustrates two radar plots for the LSDs of the sixteen phonemes under investigation as output by the three models (for 'hearing', 'speaking' and 'thinking') trained on data preprocessed with the best (lowest average mean LSD and SNR) preprocessing methods discovered (MinMax preprocessed audio and bandpass-filtered EEG with the average envelope from the Hilbert transform).

It is possible to see in Figure 12 that /m/ achieves the lowest decoded LSD across all three conditions, followed by /u:/ and /ŋ/. The best-performing model is the one decoding spoken phonemes (for all phonemes, mean LSD=2.14), followed by hearing (mean LSD=2.17) and thinking (mean LSD=2.19). Using the stacked decoder, we see an inversion of this pattern: all fricatives achieve lower decoded LSDs; however, total performance is degraded across every measure under all conditions.

The reader is encouraged to visit the FEIS database [44] to listen to the decoded audio.

Hearing	/iː/	/æ/	/uː/	/ɔː/	/f/	/v/	/s/	/z/	/ʃ/	/ʒ/	/p/	/t/	/k/	/m/	/n/	/ŋ/
BoxCox and Hilbert	25.67	25.47	25.33	25.19	26.31	25.69	26.24	25.87	26.32	25.33	25.00	25.43	25.40	25.25	25.36	25.22
	-17.13	-17.86	-16.54	-15.73	-23.45	-18.99	-22.18	-18.61	-20.47	-18.94	-19.49	-17.13	-18.10	-15.47	-16.69	-15.83
BoxCox and ICA	25.67	25.47	25.33	25.19	26.31	25.69	26.24	25.87	26.32	25.33	25.00	25.43	25.40	25.25	25.36	25.22
	-17.13	-17.86	-16.54	-15.73	-23.45	-18.99	-22.18	-18.61	-20.47	-18.94	-19.49	-17.13	-18.10	-15.47	-16.69	-15.83
BoxCox and Laplacian	25.67	25.47	25.33	25.19	26.31	25.69	26.24	25.87	26.32	25.33	25.00	25.43	25.40	25.25	25.36	25.22
	-17.13	-17.86	-16.54	-15.73	-23.45	-18.99	-22.18	-18.61	-20.47	-18.94	-19.49	-17.13	-18.10	-15.47	-16.69	-15.83
MinMax and Hilbert	1.93	2.10	1.71	1.70	2.26	2.04	2.64	2.32	2.28	2.70	2.50	2.82	2.66	**1.61**	1.74	1.71
	-0.93	-2.20	-0.87	-0.91	-2.14	-0.39	-2.72	-0.88	-2.51	-1.03	-2.37	-3.85	-2.47	-0.83	-0.92	-0.80
MinMax and ICA	2.11	2.31	1.85	1.87	2.37	2.24	2.74	2.46	2.48	2.90	2.70	3.06	2.86	1.75	1.87	1.86
	-0.69	-2.15	-0.76	-0.94	-2.33	-0.39	-2.78	-0.83	-2.42	-1.02	-3.04	-4.21	-2.90	-0.66	-0.79	-0.70
MinMax and Laplacian	2.07	2.35	1.86	1.87	2.38	2.23	2.75	2.45	2.49	2.92	2.70	3.03	2.88	1.76	1.86	1.86
	-0.75	-2.28	-0.76	-0.92	-2.45	**-0.38**	-2.85	-0.82	-2.53	-1.02	-3.13	-4.25	-2.95	-0.65	-0.71	-0.70
Z-Score and Hilbert	20.62	20.56	22.78	22.28	24.61	20.34	21.88	21.31	18.61	20.93	22.77	23.95	23.26	21.21	19.87	19.60
	-16.82	-17.75	-16.46	-15.70	-23.42	-18.75	-22.04	-18.47	-20.14	-18.78	-19.45	-17.11	-18.06	-15.29	-16.48	-15.59
Z-Score and ICA	25.67	25.47	25.33	25.19	26.31	25.69	26.24	25.05	24.80	25.33	25.00	24.93	24.58	24.55	24.64	24.33
	-17.13	-17.86	-16.54	-15.73	-23.45	-18.99	-22.18	-18.59	-20.46	-18.94	-19.49	-17.13	-18.09	-15.45	-16.67	-15.79
Z-Score and Laplacian	22.70	25.47	25.33	25.19	26.31	25.69	26.24	25.87	25.48	25.33	25.00	24.59	25.40	25.25	24.77	24.33
	-16.99	-17.86	-16.54	-15.73	-23.45	-18.99	-22.18	-18.61	-20.45	-18.94	-19.49	-17.10	-18.10	-15.47	-16.69	-15.79

Table 2: Nine permutations of models (3 audio and 3 EEG preprocessing methods) investigating the ’hearing’ condition for all 16 repeated-phoneme stimuli played to FEIS participant 19. As measured against the original audio stimuli, LSD results are presented in the shaded cells (above); SNR results are presented in the non-shaded cells (below). Best results are in bold.

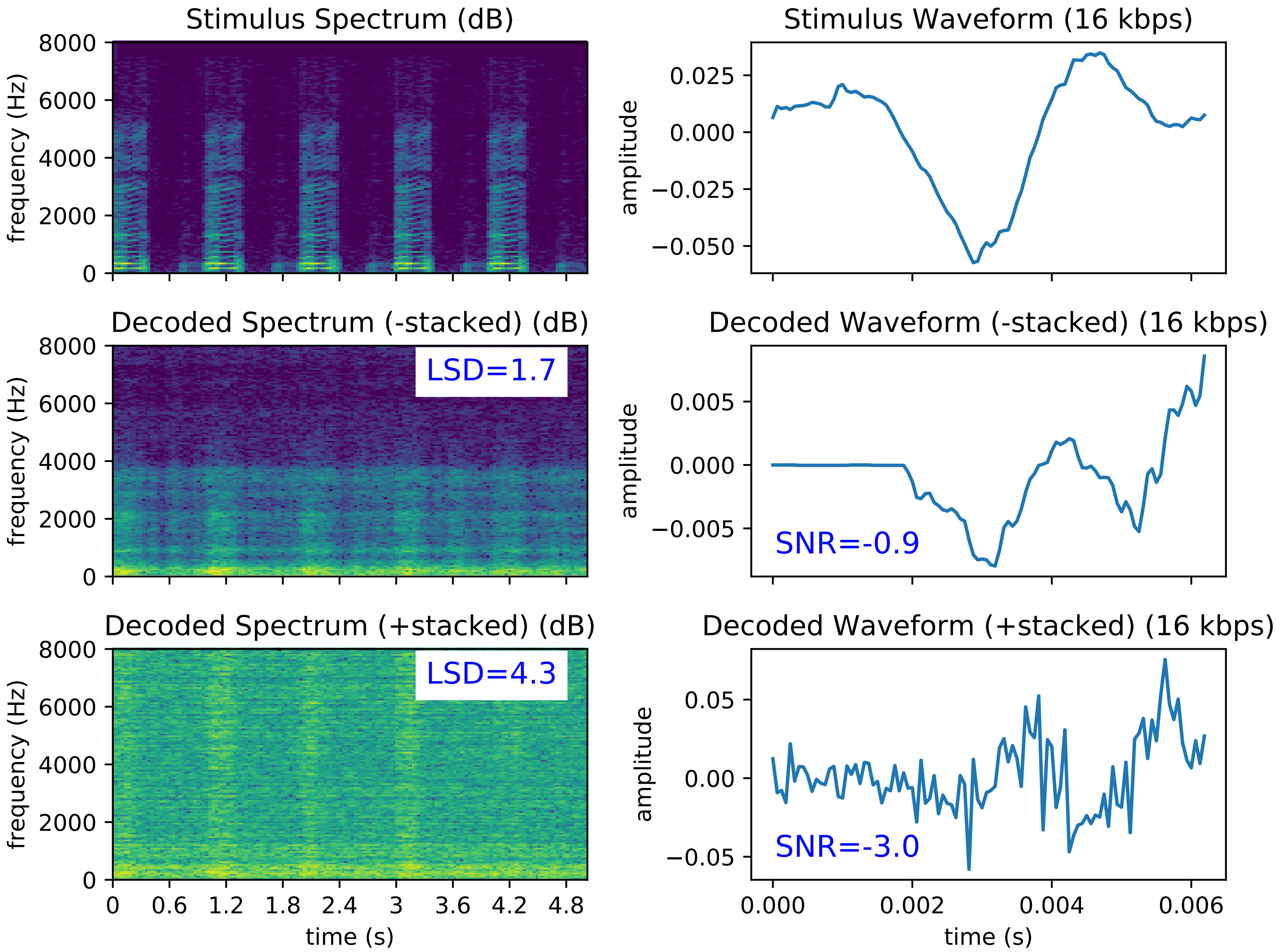


Figure 9: For the 'hearing' condition with FEIS participant 19: repeated-phoneme stimulus /m/ spectrum and waveform (top), decoded (without stacked decoder) spectrum and waveform (middle), and decoded (with stacked decoder) spectrum and waveform (bottom); Min-Max preprocessed audio and bandpass-filtered EEG with average envelope from the Hilbert transform). Code for figures adapted from [81].

Speaking	/iː/	/æ/	/uː/	/ɔː/	/f/	/v/	/s/	/z/	/ʃ/	/ʒ/	/p/	/t/	/k/	/m/	/n/	/ŋ/
BoxCox and Hilbert	25.67	25.47	25.33	25.19	26.31	25.69	26.24	25.87	26.32	25.33	25.00	25.43	25.40	25.25	25.36	25.22
	-17.13	-17.86	-16.54	-15.73	-23.45	-18.99	-22.18	-18.61	-20.47	-18.94	-19.49	-17.13	-18.10	-15.47	-16.69	-15.83
BoxCox and ICA	25.67	25.47	25.33	25.19	26.31	25.69	26.24	25.87	26.32	25.33	25.00	25.43	25.40	25.25	25.36	25.22
	-17.13	-17.86	-16.54	-15.73	-23.45	-18.99	-22.18	-18.61	-20.47	-18.94	-19.49	-17.13	-18.10	-15.47	-16.69	-15.83
BoxCox and Laplacian	25.67	25.47	25.33	25.19	26.31	25.69	26.24	25.87	26.32	25.33	25.00	25.43	25.40	25.25	25.36	25.22
	-17.13	-17.86	-16.54	-15.73	-23.45	-18.99	-22.18	-18.61	-20.47	-18.94	-19.49	-17.13	-18.10	-15.47	-16.69	-15.83
MinMax and Hilbert	1.95	2.07	**1.57**	1.70	2.27	2.06	2.58	2.33	2.34	2.64	2.45	2.74	2.64	1.63	1.62	1.61
	-1.03	-1.79	-0.93	-0.87	-2.20	-0.41	-2.82	-1.00	-2.41	-1.14	-2.25	-3.55	-2.41	-0.77	-1.03	-0.83
MinMax and ICA	2.10	2.39	1.86	1.88	2.39	2.21	2.73	2.47	2.50	2.92	2.74	3.05	2.88	1.76	1.87	1.88
	-0.63	-2.38	-0.86	-0.97	-2.39	-0.37	-2.75	-0.85	-2.48	-1.10	-3.26	-4.19	-3.00	-0.63	-0.75	-0.70
MinMax and Laplacian	2.06	2.31	1.77	1.85	2.35	2.18	2.73	2.46	2.48	2.85	2.72	2.99	2.87	1.74	1.80	1.85
	-0.60	-2.28	-0.85	-0.92	-2.32	**-0.36**	-2.83	-0.83	-2.49	-1.05	-3.12	-4.02	-2.94	-0.61	-0.69	-0.67
Z-Score and Hilbert	17.72	23.71	20.43	22.23	23.96	19.68	17.85	19.48	19.90	18.86	22.84	23.44	22.98	19.86	18.98	18.94
	-16.68	-17.82	-16.34	-15.61	-23.39	-18.74	-21.83	-18.33	-20.22	-18.65	-19.47	-17.12	-18.07	-15.28	-16.39	-15.60
Z-Score and ICA	25.67	25.47	24.57	25.19	26.31	25.69	26.24	25.87	26.32	25.33	25.00	25.43	25.40	25.25	25.36	25.22
	-17.13	-17.86	-16.51	-15.73	-23.45	-18.99	-22.18	-18.61	-20.47	-18.94	-19.49	-17.13	-18.10	-15.47	-16.69	-15.83
Z-Score and Laplacian	24.34	24.66	24.26	25.19	26.31	25.69	26.24	24.78	24.79	23.23	23.99	25.43	24.60	25.25	24.13	25.22
	-17.09	-17.85	-16.48	-15.73	-23.45	-18.99	-22.18	-18.57	-20.44	-18.87	-19.47	-17.13	-18.10	-15.47	-16.68	-15.83

Table 3: Nine permutations of models (3 audio and 3 EEG preprocessing methods) investigating the ’speaking’ condition for all 16 repeated-phoneme stimuli played to FEIS participant 19. As measured against the original audio stimuli, LSD results are presented in the shaded cells (above); SNR results are presented in the non-shaded cells (below). Best results are in bold.

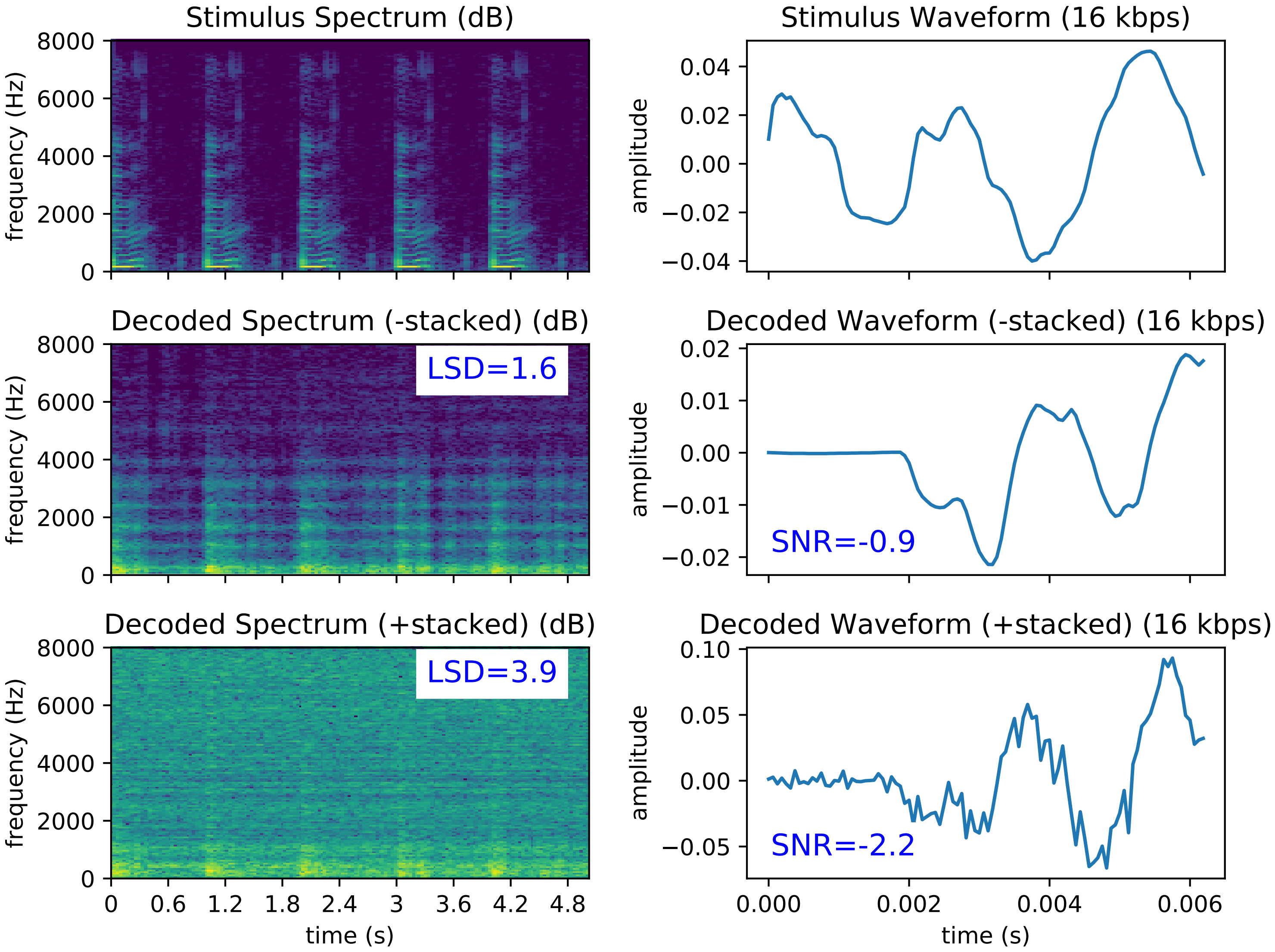


Figure 10: For the 'speaking' condition with FEIS participant 19: repeated-phoneme stimulus /u:/ spectrum and waveform (top), decoded (without stacked decoder) spectrum and waveform (middle), and decoded (with stacked decoder) spectrum and waveform (bottom); Min-Max preprocessed audio and bandpass-filtered EEG with average envelope from the Hilbert transform). Code for figures adapted from [81].

Thinking	/iː/	/æ/	/uː/	/ɔː/	/f/	/v/	/s/	/z/	/ʃ/	/ʒ/	/p/	/t/	/k/	/m/	/n/	/ŋ/
BoxCox and Hilbert	25.67	25.47	25.33	25.19	26.31	25.69	26.24	25.87	26.32	25.33	25	25.43	25.4	25.25	25.36	25.22
	-17.13	-17.86	-16.54	-15.73	-23.45	-18.99	-22.18	-18.61	-20.47	-18.94	-19.49	-17.13	-18.1	-15.47	-16.69	-15.83
BoxCox and ICA	25.67	25.47	25.33	25.19	26.31	25.69	26.24	25.87	26.32	25.33	25	25.43	25.4	25.25	25.36	25.22
	-17.13	-17.86	-16.54	-15.73	-23.45	-18.99	-22.18	-18.61	-20.47	-18.94	-19.49	-17.13	-18.1	-15.47	-16.69	-15.83
BoxCox and Laplacian	25.67	25.47	25.33	25.19	26.31	25.69	26.24	25.87	26.32	25.33	25	25.43	25.4	25.25	25.36	25.22
	-17.13	-17.86	-16.54	-15.73	-23.45	-18.99	-22.18	-18.61	-20.47	-18.94	-19.49	-17.13	-18.1	-15.47	-16.69	-15.83
MinMax and Hilbert	2.01	2.21	**1.69**	1.74	2.21	2.04	2.6	2.35	2.33	2.78	2.46	2.73	2.69	1.66	1.71	1.76
	-0.91	-2.07	-0.89	-0.91	-2.31	-0.42	-2.69	-0.94	-2.46	-1.04	-2.39	-3.48	-2.52	-0.84	-0.95	-0.74
MinMax and ICA	2.12	2.38	1.87	1.88	2.38	2.21	2.73	2.49	2.49	2.91	2.74	3.05	2.87	1.76	1.86	1.88
	-0.6	-2.42	-0.78	-0.95	-2.48	**-0.36**	-2.81	-0.88	-2.46	-1.01	-3.15	-4.17	-2.95	-0.62	-0.67	-0.65
MinMax and Laplacian	2.12	2.38	1.88	1.89	2.39	2.23	2.74	2.48	2.51	2.92	2.73	3.04	2.89	1.76	1.87	1.89
	-0.59	-2.39	-0.8	-0.96	-2.52	-0.39	-2.8	-0.84	-2.45	-1.04	-3.15	-4.13	-3.01	-0.63	-0.71	-0.65
Z-Score and Hilbert	21.34	23.7	21.95	22.11	21.7	18.51	21.36	19.67	22.62	23.68	23.58	24.59	24.75	18.93	20.39	24.65
	-16.96	-17.81	-16.43	-15.66	-23.25	-18.74	-21.97	-18.4	-20.3	-18.92	-19.47	-17.1	-18.1	-15.15	-16.53	-15.82
Z-Score and ICA	25.67	25.47	25.33	25.19	26.31	25.69	26.24	25.87	26.32	25.33	25	25.43	25.4	25.25	25.36	25.22
	-17.13	-17.86	-16.54	-15.73	-23.45	-18.99	-22.18	-18.61	-20.47	-18.94	-19.49	-17.13	-18.1	-15.47	-16.69	-15.83
Z-Score and Laplacian	25.67	25.47	25.33	25.19	26.31	25.69	26.24	25.87	26.32	25.33	25	25.43	25.4	25.25	25.36	25.22
	-17.13	-17.86	-16.54	-15.73	-23.45	-18.99	-22.18	-18.61	-20.47	-18.94	-19.49	-17.13	-18.1	-15.47	-16.69	-15.83

Table 4: Nine permutations of models (3 audio and 3 EEG preprocessing methods) investigating the ’thinking’ condition for all 16 repeated-phoneme stimuli played to FEIS participant 19. As measured against the original audio stimuli, LSD results are presented in the shaded cells (above); SNR results are presented in the non-shaded cells (below). Best results are in bold.

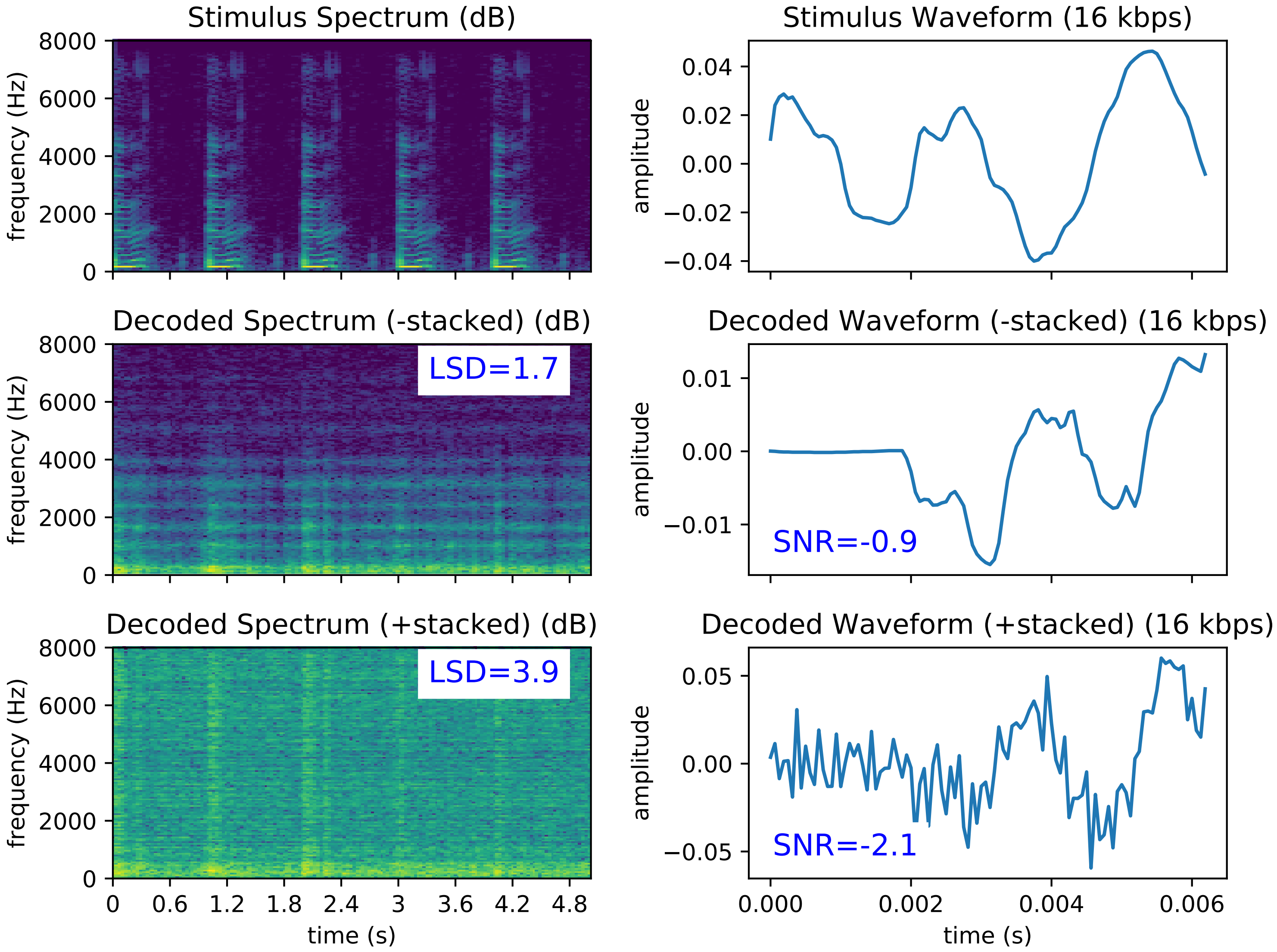


Figure 11: For the 'thinking' condition with FEIS participant 19: repeated-phoneme stimulus /u:/ spectrum and waveform (top), decoded (without stacked decoder) spectrum and waveform (middle), and decoded (with stacked decoder) spectrum and waveform (bottom); Min-Max preprocessed audio and bandpass-filtered EEG with average envelope from the Hilbert transform). Code for figures adapted from [81].

	/iː/	/æ/	/uː/	/ɔː/	/f/	/v/	/s/	/z/	/ʃ/	/ʒ/	/p/	/t/	/k/	/m/	/n/	/ŋ/
Hearing																
BoxCox	3.29	3.47	3.53	3.68	**2.73**	3.19	2.86	3.04	2.90	3.59	3.84	3.50	3.50	3.64	3.54	3.68
	-1.60	-1.99	-1.34	-0.94	-6.58	-2.77	-5.39	-2.48	-3.89	-2.72	-3.08	-1.60	-2.18	**-0.84**	-1.38	-0.99
MinMax	3.96	4.08	4.25	4.39	3.19	3.75	3.30	3.62	3.30	4.14	4.48	4.30	4.16	4.44	4.26	4.39
	-3.20	-3.25	-2.72	-2.08	-7.52	-3.85	-6.29	-3.71	-4.72	-3.83	-4.47	-3.83	-3.74	-2.37	-2.97	-2.25
Z-Score	4.44	4.54	4.80	4.94	3.67	4.21	3.76	4.07	3.74	4.58	4.93	4.60	4.57	4.94	4.76	4.93
	-4.08	-4.19	-3.77	-3.10	-9.44	-5.07	-8.19	-4.79	-6.54	-4.95	-5.54	-3.80	-4.38	-3.26	-3.82	-3.33
Speaking																
BoxCox	3.25	3.42	3.49	3.63	**2.70**	3.15	2.83	3.01	2.87	3.54	3.79	3.46	3.45	3.59	3.49	3.63
	-1.58	-1.97	-1.32	-0.93	-6.54	-2.73	-5.35	-2.45	-3.86	-2.69	-3.05	-1.58	-2.15	**-0.83**	-1.36	-0.97
MinMax	3.61	3.71	3.85	3.95	2.92	3.44	3.02	3.30	3.04	3.84	4.09	3.93	3.76	3.99	3.86	3.99
	-2.76	-2.06	-2.22	-1.42	-6.07	-2.74	-4.92	-2.95	-3.46	-2.86	-3.15	-2.81	-2.66	-1.89	-2.55	-1.77
Z-Score	4.38	4.44	4.71	4.83	3.59	4.13	3.68	3.99	3.66	4.52	4.84	4.51	4.48	4.86	4.68	4.86
	-4.00	-3.89	-3.71	-2.86	-9.14	-4.86	-7.90	-4.59	-6.26	-4.83	-5.26	-3.57	-4.16	-3.11	-3.70	-3.18
Thinking																
BoxCox	3.85	4.05	4.14	4.29	3.22	3.75	3.32	3.59	3.31	4.13	4.44	4.07	4.08	4.24	4.13	4.28
	-2.51	-3.03	-2.14	-1.62	-8.05	-3.94	-6.82	-3.62	-5.22	-3.89	-4.33	-2.51	-3.23	-1.47	-2.21	-1.69
MinMax	3.60	3.73	3.86	4.02	**3.00**	3.45	3.10	3.28	3.11	3.81	4.13	3.94	3.81	4.10	3.89	4.00
	-2.46	-2.22	-2.11	-1.52	-6.25	-2.82	-5.09	-2.81	-3.62	-2.71	-3.40	-3.04	-2.89	-1.98	-2.38	-1.62
Z-Score	3.75	3.78	4.06	4.24	**3.00**	3.52	3.11	3.38	3.12	3.88	4.17	3.87	3.83	4.28	4.03	4.16
	-2.42	-2.03	-2.21	-1.68	-6.36	-3.06	-5.19	-2.91	-3.70	-2.81	-3.07	-1.81	-2.29	-2.05	-2.11	**-1.45**

Table 5: Three stacked-decoder model variations (per audio preprocessing method) for the ’hearing’, ’speaking’, and ’thinking’ conditions, for each of the 16 repeated-phoneme stimuli played to FEIS participant 19. As measured against the original audio stimuli, LSD results are presented in the shaded cells (above); SNR results are presented in the non-shaded cells (below). Best results are in bold.

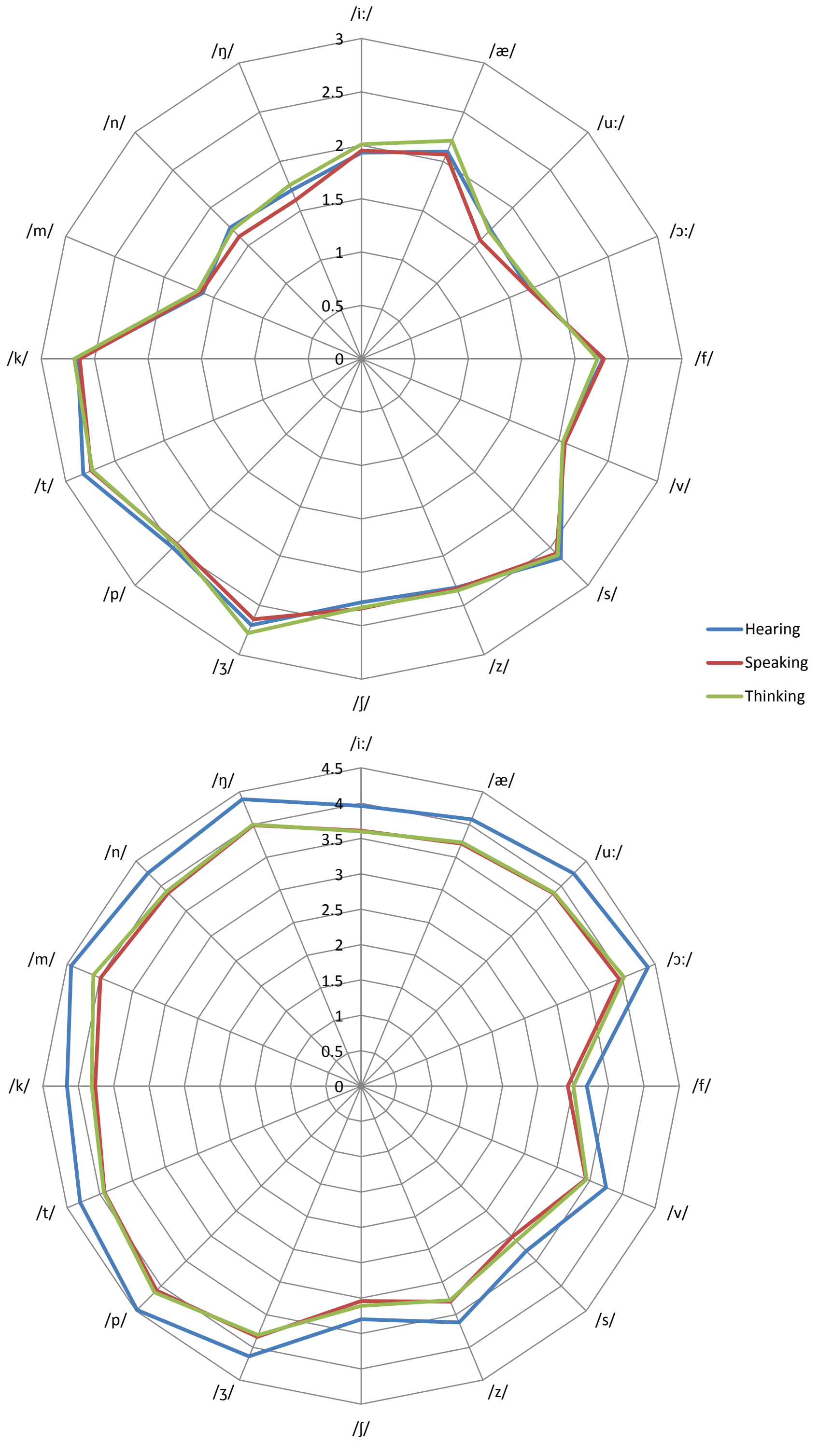


Figure 12: [Above] A radar plot of the LSDs for the sixteen phonemes under investigation, using the three models for the hearing, speaking and thinking conditions trained on the best preprocessing discovered (Min-Max audio and bandpass-filtered EEG + avg. envelope from the Hilbert transform). [Below] A radar plot of the LSDs for the same phonemes under the same conditions *after* decoding with the stacked decoder model.

6. Discussion

Our results are both surprising, and encouraging. In our hypotheses (cf. §3.4), we did not expect to achieve decoded audio output distinguishable from noise; hypothesis (1) anticipated that the (relatively) non-complex neural network architectures employed from replicating Akbari *et al.* [5] would struggle to learn from 14-channel EEG data (as opposed to Akbari *et al.*s 256-channel data).

While the results we achieved were better than noise, they were nonetheless very distorted and polluted with signal artifacts; *contra* hypothesis (3) our attempt to address this issue with a stacked decoder was unsuccessful and degraded the signal further; however, various signal-cleaning neural network architectures (such as variational or denoising autoencoders for audio preprocessing [82]) and other whitening or localisation transforms (such as wavelet analysis for EEG processing [83]) continue to be in active development for this purpose, and their inclusion into our EEG-to-audio pipeline may significantly improve decoded output in future investigations.

Hypothesis (2) was also disproved; in contrast to Zhao & Rudzicz's [40] concerns regarding severe signal distortions from articulate movement, our 'speaking' condition achieved (marginally) best regression-based decoding. However, we remain reserved with our interpretation of the results; as is clearly illustrated in Figure 12, the overall 'trend' of objective LSD measures for each phoneme remains constant under each condition (this remains true even for the noisy stacked decoder output), which may suggest phenomena at play that we had not accounted for in our experimental design.

Angrick *et al.* [84] propose that, currently, deep CNN models offer the best solution for regression tasks involving EEG data. This may be due to the non-linear mapping between the spatio-temporal features of both continuous speech and continuous EEG data; in a deep CNN, the convolutional layers are able to capitalise on (and extract higher-level features from) the relationship between channel (electrode) position and time [85]—if forced to preprocess EEG data to lower two- or one-dimensional feature arrays for, e.g., a feedforward network, this latent information becomes lost. It would be interesting, therefore, to take the CNN model architecture used within our preliminary results (§3.2; Deep4Net within the BrainDecode EEG toolbox [45]), and adpated this to a regression-based decoding task to see if results can be improved.

We note also that our results are impacted by a limited amount of data; typically, the more trials that used for epoch averaging of the EEG data, the greater the increase in SNR. Therefore the average of many trials is needed for reliable results. Luck [86] notes that for measuring large component effects (such as the those for the P300 speller) it would be typical to record 30-40 trials per condition for the benefit of SNR averaging. However, SNR is also improved by the density of the EEG device; Akbari *et al.* [5] use a medical-grade high density ECoG device of 256 electrodes—this was also an empirical motivation behind using 256-valued bottleneck features—and report good SNR with only 6 trials per condition. By this standard, we may not have enough trials per phoneme per subject, especially because the cognitive phenomena we are measuring are expected to be subtler than P300 responses. Machine learning architectures have been proposed to address the limitations of small dataset size (e.g. by generating artificial EEG signal data using generative adverserial networks [87]), and future directions for the FEIS dataset [44] may wish to explore these methods.

Much of our investigation (and computing power) addressed different data preprocessing methods; *prima facie*, Z-Score and Box-Cox Gaussianisation techniques were not suited to the WORLD vocoder features; while we anticipated that one method would be better suited to the task, we did not expect results to be universally worse than unprocessed, raw data (Table 1). Were the models retrained using MFCCs instead of WORLD [49] features, we would wish to investigate whether this unexpected result is retained when preprocessing the data using, for instance, cepstral mean normalization (CMN); moreover, this would also allow us to investigate if an MFCC-trained autoencoder can achieve comparable decoding results, and whether the failure of our stacked decoder model was the product of non-optimized implementation or difficulty to learn from vocoder features with severe noise distortion.

7. Conclusion

In this investigation, we have discovered that it is possible to achieve discernible, decoded single-phoneme audio output from a low-density EEG headset using the EEG signal data alone for heard, spoken and imagined phonemes. However, the ’degree’ of discernibly is in question: as regards intelligibility, it is easy to fool oneself into hearing the phoneme one is ’expecting’ to hear, and a listening test with naïve participants would allow an objective assessment.

Different preprocessing techniques have demonstrated that this is a very important step to ’get right’; it is possible to have normalised or standardised data that affords excellent learning for a neural network at one end of the pipeline (audio or EEG), but preprocessed data that the network struggles to learn from at the *other* end of the pipeline will lead to polluted or distorted data. It is a game of two halves, both of which must be played correctly or not at all.

The ’next step’ for subsequent investigations is to investigate the feasibility of decoding continuous speech using a low-density EEG device. Researchers are making impressive progress in this regard using high-density, medical-grade EEG devices (cf. §2.3); however, with the ultimate goal of speech neuroprosthesis, it is our desire to take the machine learning architectures and data processing techniques from those researchers and *scale* them to achieve comparable results on a commercially-available device, making such AACs available to all.

Bibliography

[1] J. Clayton, "Towards phone classification from imagined speech using a lightweight EEG brain-computer interface," University of Edinburgh, Edinburgh, UK, 2019, [unpublished M.Sc. dissertation].

[2] M. Angrick, C. Herff, E. Mugler, M. C. Tate, M. W. Slutzky, D. J. Krusienski, and T. Schultz, "Speech synthesis from ECoG using densely connected 3D convolutional neural networks," *Journal of neural engineering*, vol. 16, no. 3, p. 036019, 2019.

[3] G. K. Anumanchipalli, J. Chartier, and E. F. Chang, "Speech synthesis from neural decoding of spoken sentences," *Nature*, vol. 568, no. 7753, p. 493, 2019.

[4] P. Saha and S. Fels, "Hierarchical deep feature learning for decoding imagined speech from EEG," *arXiv preprint arXiv:1904.04352*, 2019.

[5] H. Akbari, B. Khalighinejad, J. L. Herrero, A. D. Mehta, and N. Mesgarani, "Towards reconstructing intelligible speech from the human auditory cortex," *Scientific reports*, vol. 9, no. 1, p. 874, 2019.

[6] M. J. Vansteensel, E. G. Pels, M. G. Bleichner, M. P. Branco, T. Denison, Z. V. Freudenburg, P. Gosselaar, S. Leinders, T. H. Ottens, M. A. Van Den Boom *et al.*, "Fully implanted brain–computer interface in a locked-in patient with ALS," *New England Journal of Medicine*, vol. 375, no. 21, pp. 2060–2066, 2016.

[7] E. Musk and Neurolink, "An integrated brain-machine interface platform with thousands of channels," *bioRxiv*, 2019. [Online]. Available: https://www.biorxiv.org/content/early/2019/07/18/703801 [Accessed 12/09/2019]

[8] Emotiv EPOC+. [Online]. Available: https://www.emotiv.com/epoc/ [Accessed

03/09/2019]. doi: https://doi.org/10.5281/zenodo.3369179

[9] Y. S. Su, A. Veeravagu, and G. Grant, “Neuroplasticity after traumatic brain injury,” *Translational Research in Traumatic Brain Injury, Laskowitz D, Grant G, editors. Boca Raton (FL): CRC Press/Taylor and Francis Group, USA*, 2016.

[10] A. M. Chilosi, P. Cipriani, B. Bertuccelli, L. Pfanner, and G. Cioni, “Early cognitive and communication development in children with focal brain lesions,” *Journal of Child Neurology*, vol. 16, no. 5, pp. 309–316, 2001.

[11] M. Altinay, E. Estemalik, and D. A. Malone Jr, “A comprehensive review of the use of deep brain stimulation (DBS) in treatment of psychiatric and headache disorders,” *Headache: The Journal of Head and Face Pain*, vol. 55, no. 2, pp. 345–350, 2015.

[12] S. Lowel and W. Singer, “Selection of intrinsic horizontal connections in the visual cortex by correlated neuronal activity,” *Science*, vol. 255, no. 5041, pp. 209–212, 1992.

[13] F. Babiloni, F. Cincotti, F. Carducci, P. Rossini, and C. Babiloni, “Spatial enhancement of EEG data by surface laplacian estimation: The use of magnetic resonance imaging-based head models,” *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology*, vol. 112, pp. 724–7, 06 2001.

[14] B. Burle, L. Spieser, C. Roger, L. Casini, T. Hasbroucq, and F. Vidal, “Spatial and temporal resolutions of EEG: Is it really black and white? A scalp current density view,” *International Journal of Psychophysiology*, vol. 97, no. 3, pp. 210–220, 2015.

[15] M. D. Nunez, P. L. Nunez, R. Srinivasan, H. Ombao, M. Linquist, W. Thompson, and J. Aston, “Electroencephalography (EEG): neurophysics, experimental methods, and signal processing,” *Handbook of Neuroimaging Data Analysis*, pp. 175–197, 2016.

[16] J. W. Britton, L. C. Frey, J. Hopp, P. Korb, M. Koubeissi, W. Lievens, E. Pestana-Knight, and E. L. St, *Electroencephalography (EEG): An introductory text and atlas of normal and abnormal findings in adults, children, and infants*. American Epilepsy Society, Chicago, 2016.

[17] D. Hagemann and E. Naumann, “The effects of ocular artifacts on (lateralized) broadband power in the EEG,” *Clinical Neurophysiology*, vol. 112, no. 2, pp. 215–231, 2001.

[18] S. Sreeja, R. R. Sahay, D. Samanta, and P. Mitra, "Removal of eye blink artifacts from EEG signals using sparsity," *IEEE journal of biomedical and health informatics*, vol. 22, no. 5, pp. 1362–1372, 2017.

[19] X. Lei and K. Liao, "Understanding the influences of EEG reference: A large-scale brain network perspective," *Frontiers in Neuroscience*, vol. 11, p. 205, 2017. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnins.2017.00205, doi: 10.3389/fnins.2017.00205

[20] G. A. Light, L. E. Williams, F. Minow, J. Sprock, A. Rissling, R. Sharp, N. R. Swerdlow, and D. L. Braff, "Electroencephalography (EEG) and event-related potentials (ERPs) with human participants," *Current protocols in neuroscience*, vol. 52, no. 1, pp. 6–25, 2010.

[21] S. P. Ahlfors, J. Han, J. W. Belliveau, and M. S. Hämäläinen, "Sensitivity of MEG and EEG to source orientation," *Brain topography*, vol. 23, no. 3, pp. 227–232, 2010.

[22] M. De Vos, M. Kroesen, R. Emkes, and S. Debener, "P300 speller BCI with a mobile EEG system: comparison to a traditional amplifier," *Journal of neural engineering*, vol. 11, no. 3, p. 036008, 2014.

[23] G. Hickok and D. Poeppel, "The cortical organization of speech processing," *Nature reviews neuroscience*, vol. 8, no. 5, p. 393, 2007.

[24] M. Vigneau, V. Beaucousin, P.-Y. Herve, H. Duffau, F. Crivello, O. Houde, B. Mazoyer, and N. Tzourio-Mazoyer, "Meta-analyzing left hemisphere language areas: phonology, semantics, and sentence processing," *Neuroimage*, vol. 30, no. 4, pp. 1414–1432, 2006.

[25] J. Pujol, J. Deus, J. M. Losilla, and A. Capdevila, "Cerebral lateralization of language in normal left-handed people studied by functional MRI," *Neurology*, vol. 52, no. 5, pp. 1038–1038, 1999.

[26] M. Jefferson, "Usability of automatic speech recognition systems for individuals with speech disorders: past, present, future, and a proposed model," 2019, the University of Minnesota Digital Conservancy, http://hdl.handle.net/11299/20275.

[27] E. Yılmaz, V. Mitra, G. Sivaraman, and H. Franco, "Articulatory and bottleneck features for speaker-independent ASR of dysarthric speech," *Computer Speech & Language*, vol. 58, pp. 319–334, 2019.

[28] L. A. Moctezuma, A. A. Torres-García, L. Villaseñor-Pineda, and M. Carrillo,

"Subjects identification using EEG-recorded imagined speech," *Expert Systems with Applications*, vol. 118, pp. 201–208, 2019.

[29] B. U. Forstmann, E.-J. Wagenmakers *et al.*, *An introduction to model-based cognitive neuroscience*. Springer, 2015.

[30] P. A. Abhang, B. W. Gawali, and S. C. Mehrotra, *Introduction to EEG-and speech-based emotion recognition*. Academic Press, 2016.

[31] M. Duvinage, T. Castermans, T. Dutoit, M. Petieau, T. Hoellinger, C. D. Saedeleer, K. Seetharaman, and G. Cheron, "A P300-based quantitative comparison between the Emotiv EPOC headset and a medical EEG device," *Biomedical Engineering*, vol. 765, no. 1, pp. 2012–2764, 2012.

[32] D. S. Benitez, S. Toscano, and A. Silva, "On the use of the Emotiv EPOC neuroheadset as a low cost alternative for EEG signal acquisition," in *2016 IEEE Colombian Conference on Communications and Computing (COLCOM)*. IEEE, 2016, pp. 1–6.

[33] A. R. Sereshkeh, R. Yousefi, A. T. Wong, and T. Chau, "Online classification of imagined speech using functional near-infrared spectroscopy signals," *Journal of neural engineering*, vol. 16, no. 1, p. 016005, 2018.

[34] C. Cooney, F. Raffaella, and D. Coyle, "Classification of imagined spoken word-pairs using convolutional neural networks," in *The 8th Graz BCI Conference, 2019*, 2019.

[35] S. Martin, I. Iturrate, P. Brunner, J. d. R. Millán, G. Schalk, R. T. Knight, and B. N. Pasley, "Individual word classification during imagined speech using intracranial recordings," in *Brain-Computer Interface Research*. Springer, 2019, pp. 83–91.

[36] S. Martin, I. Iturrate, J. d. R. Millán, R. T. Knight, and B. N. Pasley, "Decoding inner speech using electrocorticography: progress and challenges toward a speech prosthesis," *Frontiers in neuroscience*, vol. 12, p. 422, 2018.

[37] C. Herff, D. Heger, A. De Pesters, D. Telaar, P. Brunner, G. Schalk, and T. Schultz, "Brain-to-text: decoding spoken phrases from phone representations in the brain," *Frontiers in neuroscience*, vol. 9, p. 217, 2015.

[38] G. Krishna, Y. Han, C. Tran, M. Carnahan, and A. H. Tewfik, "State-of-the-art speech recognition using EEG and towards decoding of speech spectrum from EEG," *arXiv preprint arXiv:1908.05743*, 2019.

[39] H. Watanabe, H. Tanaka, S. Sakti, and S. Nakamura, “Synchronization between overt speech envelope and EEG oscillations during imagined speech,” *Neuroscience research*, 2019.

[40] S. Zhao and F. Rudzicz, “Classifying phonological categories in imagined and articulated speech,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 992–996.

[41] M. D’Zmura, S. Deng, T. Lappas, S. Thorpe, and R. Srinivasan, “Toward EEG sensing of imagined speech,” in *Human-Computer Interaction. New Trends*. Springer Berlin Heidelberg, 2009, pp. 40–48.

[42] S. Vela and L. Carlos, “Reconocimiento del habla silenciosa con seales electroencefalogrficas (EEG) para interfaces cerebro-computador,” 2015, doctoral thesis, Universidad Nacional de Colombia - Sede Bogot.

[43] A. P. Nguyen CH, Karavas GK, “Inferring imagined speech using EEG signals: a new approach using Riemannian manifold features,” *Journal of Neural Engineering*, vol. 15(1), p. 016002, 02 2018.

[44] S. Wellington and J. Clayton, “Fourteen-channel EEG with imagined speech (FEIS) dataset,” Aug 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3369179

[45] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, “Deep learning with convolutional neural networks for EEG decoding and visualization,” *Human brain mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[46] A. Gramfort, M. Luessi, E. Larson, D. A. Engemann, D. Strohmeier, C. Brodbeck, L. Parkkonen, and M. S. Hämäläinen, “MNE software for processing MEG and EEG data,” *Neuroimage*, vol. 86, pp. 446–460, 2014.

[47] A. Gramfort, M. Luessi, E. Larson, D. A. Engemann, D. Strohmeier, C. Brodbeck, R. Goj, M. Jas, T. Brooks, L. Parkkonen *et al.*, “MEG and EEG data analysis with MNE-Python,” *Frontiers in neuroscience*, vol. 7, p. 267, 2013.

[48] L. Sun, Y. Liu, and P. J. Beadle, “Independent component analysis of EEG signals,” in *Proceedings of 2005 IEEE International Workshop on VLSI Design and Video Technology, 2005.*, May 2005, pp. 219–222.

[49] M. Morise, F. Yokomori, and K. Ozawa, “WORLD: a vocoder-based high-quality speech synthesis system for real-time applications,” *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[50] Y. Renard, F. Lotte, G. Gibert, M. Congedo, E. Maby, V. Delannoy, O. Bertrand, and A. Lécuyer, "Openvibe: An open-source software platform to design, test, and use brain–computer interfaces in real and virtual environments," *Presence: teleoperators and virtual environments*, vol. 19, no. 1, pp. 35–53, 2010.

[51] L. Thaler, A. C. Schütz, M. A. Goodale, and K. R. Gegenfurtner, "What is the best fixation target? the effect of target shape on stability of fixational eye movements," *Vision Research*, vol. 76, pp. 31–42, 2013.

[52] M. Traxler and M. A. Gernsbacher, *Handbook of psycholinguistics*. Elsevier, 2011.

[53] C. Phillips, T. Pellathy, A. Marantz, E. Yellin, K. Wexler, D. Poeppel, M. McGinnis, and T. Roberts, "Auditory cortex accesses phonological categories: an MEG mismatch study," *Journal of Cognitive Neuroscience*, vol. 12, no. 6, pp. 1038–1055, 2000.

[54] SoX: Sound eXchange. [Online]. Available: http://sox.sourceforge.net/ [Accessed 03/09/2019]

[55] CymatiCorp, "Cykit 3.0," 2019. [Online]. Available: https://github.com/CymatiCorp/CyKit [Accessed 12/09/2019]

[56] S. Takaki and J. Yamagishi, "A deep auto-encoder based low-dimensional feature extraction from FFT spectral envelopes for statistical parametric speech synthesis," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5535–5539.

[57] H. Kawahara, "STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds," *Acoustical science and technology*, vol. 27, no. 6, pp. 349–353, 2006.

[58] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *NIPS Autodiff Workshop*, 2017.

[59] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 357–362.

[60] C. Veaux, J. Yamagishi, and K. MacDonald, "CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit, [sound]. University of Edinburgh. The Centre for Speech Technology Research (CSTR)," 2017. [Online]. Available: https://doi.org/10.7488/ds/1994

[61] O. Watts, “hands-on_tts/world_features.py,” 2019. [Online]. Available: https://github.com/oliverwatts/hands-on_tts [Accessed 12/09/2019]

[62] J. Hsu, “Pyworldvocoder - a Python wrapper for WORLD vocoder,” 2019. [Online]. Available: https://github.com/JeremyCCHsu/Python-Wrapper-for-World-Vocoder [Accessed 12/09/2019]

[63] Z. Jin, A. Finkelstein, G. J. Mysore, and J. Lu, “FFTNet: A real-time speaker-dependent neural vocoder,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2018, pp. 2251–2255.

[64] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[65] Edinburgh Compute and Data Facility, “Eddie Mark 3 compute cluster, University of Edinburgh,” 2015. [Online]. Available: www.ecdf.ed.ac.uk [Accessed 12/09/2019]

[66] K. Choi, G. Fazekas, M. Sandler, and K. Cho, “A comparison of audio signal preprocessing methods for deep neural networks on music tagging,” in *2018 26th European Signal Processing Conference (EUSIPCO).* IEEE, 2018, pp. 1870–1874.

[67] W. A. Sethares, *Rhythm and transforms.* Springer Science & Business Media, 2007.

[68] A. S. Al-Fahoum and A. A. Al-Fraihat, “Methods of EEG signal features extraction using linear analysis in frequency and time-frequency domains,” *ISRN neuroscience*, vol. 2014, 2014.

[69] I. A. Basheer and M. Hajmeer, “Artificial neural networks: fundamentals, computing, design, and application,” *Journal of microbiological methods*, vol. 43, no. 1, pp. 3–31, 2000.

[70] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, “Data preprocessing for supervised leaning,” *International Journal of Computer Science*, vol. 1, no. 2, pp. 111–117, 2006.

[71] M. Bicego and S. Baldo, “Properties of the Box–Cox transformation for pattern classification,” *Neurocomputing*, vol. 218, pp. 390–400, 2016.

[72] E. Jones, T. Oliphant, P. Peterson *et al.*, “SciPy: Open source scientific tools for Python,” 2001–2019. [Online]. Available: http://www.scipy.org/ [Accessed 12/09/2019]

[73] F. T. Smulders, S. Ten Oever, F. C. Donkers, C. W. Quaedflieg, and V. van de Ven, “Single-trial log transformation is optimal in frequency analysis of resting EEG alpha,” *European Journal of Neuroscience*, vol. 48, no. 7, pp. 2585–2598, 2018.

[74] O. Etard and T. Reichenbach, “Neural speech tracking in the theta and in the delta frequency band differentially encode clarity and comprehension of speech in noise,” *Journal of Neuroscience*, vol. 39, no. 29, pp. 5750–5759, 2019.

[75] A. Keitel, J. Gross, and C. Kayser, “Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features,” *PLoS biology*, vol. 16, no. 3, p. e2004473, 2018.

[76] T.-P. Jung, S. Makeig, C. Humphries, T.-W. Lee, M. J. Mckeown, V. Iragui, and T. J. Sejnowski, “Removing electroencephalographic artifacts by blind source separation,” *Psychophysiology*, vol. 37, no. 2, pp. 163–178, 2000.

[77] T. M. Lau, J. T. Gwin, and D. P. Ferris, “How many electrodes are really needed for EEG-based mobile brain imaging?” *Journal of Behavioral and Brain Science*, vol. 2, no. 03, p. 387, 2012.

[78] C. Carvalhaes and J. A. de Barros, “The surface Laplacian technique in EEG: Theory and methods,” *International Journal of Psychophysiology*, vol. 97, no. 3, pp. 174–188, 2015.

[79] J. Kayser and C. E. Tenke, “Issues and considerations for using the scalp surface Laplacian in EEG/ERP research: a tutorial review,” *International Journal of Psychophysiology*, vol. 97, no. 3, pp. 189–209, 2015.

[80] A. Prodeus, V. Didkovskyi, M. Didkovska, I. Kotvytskyi, D. Motorniuk, and A. Khrapachevskyi, “Objective and subjective assessment of the quality and intelligibility of noised speech,” in *2018 International Scientific-Practical Conference Problems of Infocommunications. Science and Technology (PIC S&T)*. IEEE, 2018, pp. 71–74.

[81] J. Heatherly, “EnglishSpeechUpsampler,” 2017. [Online]. Available: https://github.com/jhetherly/EnglishSpeechUpsampler.git [Accessed 12/09/2019]

[82] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, “Voice conversion from non-parallel corpora using variational auto-encoder,” in *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE, 2016, pp. 1–6.

[83] M. Mamun, M. Al-Kadi, and M. Marufuzzaman, “Effectiveness of wavelet denoising on electroencephalogram signals,” *Journal of applied research and technology*, vol. 11, no. 1, pp. 156–160, 2013.

[84] M. Angrick, C. Herff, G. Johnson, J. Shih, D. Krusienski, and T. Schultz, “Interpretation of convolutional neural networks for speech spectrogram regression from intracranial recordings,” *Neurocomputing*, vol. 342, pp. 145–151, 2019.

[85] F. A. Heilmeyer, R. Schirrmeister, L. Fiederer, M. Volker, J. Behncke, and T. Ball, “A large-scale evaluation framework for EEG deep learning architectures,” in *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 10 2018, pp. 1039–1045, doi: 10.1109/SMC.2018.00 185.

[86] S. J. Luck, *An introduction to the event-related potential technique*. MIT press, 2014, pp. 262–263.

[87] Q. Zhang and Y. Liu, “Improving brain computer interface performance by data augmentation with conditional deep convolutional generative adversarial networks,” *arXiv preprint arXiv:1806.07108*, 2018.