Human-Machine Cognitive Coalescence through an Internal Duplex Interface

by

Arnav Kapur

Submitted to the Program in Media Arts and Sciences in partial fulfillment of the requirements for the degree of Master of Science in Media Arts and Sciences at the MASSACHUSETTS INSTITUTE OF TECHNOLOGY September 2018

© Massachusetts Institute of Technology 2018. All rights reserved.

Signature redacted

Author .........................

Program in Media Arts and Sciences

August 10, 2018

Signature redacted

Certified by ....................

Pattie Maes
Professor of Media Arts and Sciences
Thesis Supervisor

Signature redacted

Accepted by ...............

Tod Machover
Academic Head, Program in Media Arts and Sciences
Human-Machine Cognitive Coalescence through an Internal Duplex Interface

Arnav Kapur

Signature redacted

Pattie Maes ......................
Thesis Supervisor
Professor of Media Arts and Sciences
MIT Media Lab

Signature redacted

Edward S. Boydier ..............
Thesis Reader
Associate Professor of Media Arts and Sciences
MIT Media Lab

Signature redacted

Antonio Torralba ..............
Thesis Reader
Professor of Electrical Engineering and Computer Science
MIT Computer Science and Artificial Intelligence Lab
Human-Machine Cognitive Coalescence through an Internal Duplex Interface

by

Arnav Kapur

Submitted to the Program in Media Arts and Sciences on September 4, 2018, in partial fulfillment of the requirements for the degree of Master of Science in Media Arts and Sciences

Abstract

In this thesis, we present a non-invasive and non-intrusive system that enables silent duplex human-machine communication and enables an interface that is internal to the user. We present a peripheral nerve-computer interface, AlterEgo, that allows a user to silently converse with a computing device without any voice or any discernible movements - thereby enabling the user to communicate with devices, AI assistants, applications or other people in a silent, concealed and seamless manner. A user’s volitional internally articulated speech is characterized by efferent signal signatures in internal speech articulators that are captured and recognized by the proposed system. The hope is to facilitate a natural language user interface, where users can silently communicate in natural language and receive information and sensory input aurally through bone conduction. This enables a discreet, closed-loop interface with a computing device, and thus providing a seamless form of cognitive augmentation. The goal of the thesis is to describe the architecture, design, implementation and operation of the entire system along with demonstrating the utility of the platform as a personal computing system.

Thesis Supervisor: Pattie Maes
Title: Professor of Media Arts and Sciences
Acknowledgments

To Pattie Maes, thank you for being an incredible advisor, researcher and person. Thank you for your unwavering support, patience and guidance throughout my stay here at MIT. Thank you for trusting me in a playground. I will always cherish your mentorship. Thank you for recognizing something in me and bringing me home.

To Antonio Torralba and Ed Boyden for their patience and feedback. Thank you Antonio for your enthusiasm and the office discussions. Thank you Ed for your feedback on formulation and class conversations.

To Keira Horowitz, thank you for the work that you do. Thank you for your patience and support.

To Harshit Agrawal, thank you for your friendship and for the best times.

To Kshitij Marwah, thank you for your friendship and all the elder brother wisdom I could have never gotten otherwise.

To Mina, thank you for being my rock, for your compassion, for making the days brighter and for your unconditional support.

To Shreyas, thank you for being my best friend, for being the most precious reminder and for everything else.

To Papa, thank you for your trust, unconditional support and for everything else.

To Ma, thank you for being you and for your belief, thank you for your unconditional everything. I can never thank you enough.
Human-Machine Cognitive Coalescence through an Internal Duplex Interface

Arnav Kapur

Pattie Maes .................................................................
Thesis Supervisor
Professor of Media Arts and Sciences
MIT Media Lab

Edward S. Boyden .......................................................1
Thesis Reader
Associate Professor of Media Arts and Sciences
MIT Media Lab

Antonio Torralba ........................................................
Thesis Reader
Professor of Electrical Engineering and Computer Science
MIT Computer Science and Artificial Intelligence Lab
Contents

1 Introduction ................................................................. 17
   1.1 Contributions .................................................... 18
   1.2 Evaluation of AlterEgo ....................................... 19
   1.3 Thesis Roadmap .................................................. 19

2 Related Work .............................................................. 21
   2.1 Invasive Systems ............................................... 21
   2.2 Non-Invasive Systems ......................................... 22

3 Methods and Implementation ......................................... 25
   3.1 Neural Control of Speech .................................... 25
   3.2 Internal Articulation .......................................... 27
   3.3 Signal Conditioning and Processing .......................... 29
   3.4 Indiscernible Silent Speech Recognition .................. 31
   3.5 Closed-Loop Interface ...................................... 34

4 Evaluation ................................................................. 39
   4.1 Silent Speech Recognition .................................. 39
   4.2 Information Rate .............................................. 42

5 Application Examples .................................................. 47
   5.1 Closed-Loop Computing ..................................... 47
   5.2 Open-Loop Computing ....................................... 50
   5.3 Telecommunications .......................................... 50
6 Future Work

6.1 Vocabulary Recognition ........................................... 53
6.2 Physiological Rehabilitation and Prosthesis .......................... 54
6.3 Ambulatory Conditions and Testing .................................... 55
6.4 Suprasegmental Elements ............................................. 55
6.5 User Interface Design .................................................. 55

A

A.1 Change in Information Transfer Rate ................................. 57
List of Figures

1-1 **When Computers were Human.** The origin of the word computer can be dated back to early 17th century [2][1], which meant "one who computes" and referred to a person who performed mathematical calculations. The image shows women human computers at the Jet Propulsion Lab at NASA in the 1930s. .......................... 17

3-1 **Simplified Outline showing arrangement of Corticonuclear Fiber Pathways and Cranial Nerves.** ................................. 26

3-2 **Volumetric Conduction of Signal.** The figure shows a simplified model of an efferent neural signal being conducted through tissue and measured at the electrode. In the model the voltage attenuates as $V \propto \frac{1}{r}$ which gives an intuition of voltage signal attenuation when measuring electrode and signal source are not in close proximity. ... 28

3-3 **Electrode Target Positions and Associated Muscle Regions.** It is important to note that the facial muscles are themselves neurally inactive, and are shown here to clarify positioning of electrodes. These positions were used in our human subject experiments. ............. 29

3-4 **Electrode Target Positions in Second Iteration of System.** We designed a second iteration of the wearable that would be more feasible to use in quotidian settings. The figure shows the electrode positions in the wearable. .................................................. 30

3-5 **Architecture of Spatiotemporal Convolutional Neural Network detecting Internal Articulated Speech in Real-Time.** ... 31
3-6 Architecture of Convolutional Neural Network classifying Internally Articulated Speech into Word Labels. 32

3-7 Sensorimotor Feedback during Speech Articulation. This shows the schematic of the Directions into Velocities of Articulators model of speech acquisition and production (from Guenther et al. 2006, Cortical interactions underlying the production of speech sounds [16]). In the silent speech interface, speech is internally articulated without phonation and sensory feedback is given through bone conduction. 33

3-8 Hardware Arrangement of the AlterEgo Wearable. 33

3-9 First Iteration of the AlterEgo Prototype. 35

3-10 Second Iteration of the AlterEgo Prototype. 36

3-11 Flowchart of the Silent Speech Interface. The figure shows the various steps of the system. The step when various features are calculated for feature fusion is an alternate strategy, which feeds computed features into the network. 37

4-1 Word Accuracy of Silent Speech Recognition. The model is evaluated on 10 users, using 10 runs of cross-validation of the dataset with arithmetic digits (0-9). The edges of boxes represent 1st and 3rd quartile respectively, and the whiskers extend to 100% coverage. 41

4-2 Online System measuring Information Transfer Rate. A visual cue of a word is displayed on screen. The rising edge of voltage coincides with the start of the subject internally articulating the word through a push-down button switch and held through the duration of articulation. The falling edge coincides with the end of articulation. The raw data during the period of high signal ($T$) is then further processed for recognition. 43
5-1 **Coalesced Human-AI Decision Making in the game of Chess.**
The figure shows a user internally communicating the previous move played by an opponent, to a game playing AI agent running in the background, and seeks counsel on next move to be played. The AI agent optimizes for the next move and relays the information to the user, through bone-conduction.

5-2 **Controlling IoT Devices through AlterEgo by Internally Articulating Commands.** The figure shows the user (left-right) internally articulating the phrases ”bulb 1 2 on”. The system recognizes the commands, sending wireless instructions to an Internet enabled switch, that switches on bulb 1 and bulb 2. This example demonstrates how remote control becomes endogenous to the user.
List of Tables

3.1 Motor and Sensory Functions of Cranial Nerves relevant to Speech and Hearing. The primary efferent signaling, during internal articulation, is through CN IX, CN X, CN XII. .......................... 27

4.1 Information Transfer Rate of AlterEgo. The table shows the average ITR (in bps) of the system, with detection of 10 words ($N = 10$). An average ITR was computed for each participant over 10 experiment trials for each participant. The table shows average trial duration (in ms) for each participant. .......................... 44

4.2 Information Transfer Rates of Brain- and Peripheral Nerve-Computer Interfaces. The table lists previously reported ITR across a range of systems. .......................... 44
Chapter 1

Introduction

Figure 1-1: When Computers were Human. The origin of the word computer can be dated back to early 17th century [2][1], which meant "one who computes" and referred to a person who performed mathematical calculations. The image shows women human computers at the Jet Propulsion Lab at NASA in the 1930s.

The origin of personal computing has its roots in the ambition of augmenting human cognition [10] [24] [3]. Since then, input methodologies to computers have come a long way since punchcards and present day input modalities have enabled computing devices to become an intrinsic part of our lives. Keyboards replaced punch cards to
facilitate text input on early computers. The modern age of mobile and ubiquitous computing ushered in the widespread adoption of voice input for communication and search applications. Natural user interfaces (NUI) including gesture-based inputs, touch and voice have been touted as natural extensions of the human persona [5] [17]. Despite significant advances made, machines, input modalities and interactivity still exist as external experiences to the human user, representing an obstacle to fully realize integration of human and machine intelligence.

We present a strategy for enabling a closed-loop non-invasive and non-intrusive human-machine interface with the subjective experience of the interaction being completely endogenous to the human user. The system is instantiated as a wearable peripheral nerve-computer interface, named AlterEgo [23], that allows a user to provide arbitrary text input to a computing device or other people using natural language, without observable action and without voice. This is done during volitional activation of internal speech articulators during internal articulation of words, a highly attenuated manner of speech. The volitional motor action during internal articulation does not engage observable speech articulators and does not require phonation which renders the interface soundless and private. Internal articulation of speech is concomitant with a subjective experience of inner speech of the user. The efferent neural signaling during motor activation is captured and processed as input, thereby giving the user volitional control over input to a computer. The feedback from a machine is transmitted to the user aurally through bone conduction transducers, that provide information and sensory feedback otherwise experienced during voiced speech. This allows a bi-directional human-machine communication in natural language.

1.1 Contributions

We make the following contributions:

- We introduce a novel model for a closed-loop silent speech system, that allows the human-computer interaction to be internal to the user, where the user receives aural information and sensory feedback without any observable action.
• We are the first to demonstrate a peripheral nerve-computer interface that enables conversational bi-directional communication with the computer, in natural language, at communication speeds resembling the speeds of regular voiced speech.

• We outline the process of internal articulation of speech and neuromuscular input needed to enable such a non-invasive and non-intrusive interface. We present the methodology needed to measure neural activation of supralaryngeal muscles, during the interaction.

• We have instantiated the system as a wireless wearable platform, AlterEgo, that enables usage in a non-clinical setting. We demonstrate the utility of the system as a personal computing platform.

1.2 Evaluation of AlterEgo

We built a prototype of AlterEgo and evaluated the system during human subject experiments. We conducted experiments to evaluate the accuracy of silent speech word recognition. These were conducted with human subjects in a stationary setting who had no history of neurological or speech related pathology. The user would internally articulate a phrase in a non-continuous manner to segregate each word for signal classification. In a 10 person study, we report a mean accuracy of 92.01% for a 10 word vocabulary. Secondly, we sought to evaluate the information transfer rate (ITR) of such a communication channel. In similar experimental conditions, we report an average ITR of 10.302 bps.

1.3 Thesis Roadmap

The thesis is organized as follows. Chapter 2 covers related work, followed by Chapter 3 which outlines the system design, methodology and implementation. Chapter 4 presents the evaluation of the AlterEgo system. Chapter 5 presents potential example
applications of such an interface. Finally, Chapter 6 presents planned future work.
Chapter 2

Related Work

The interest in integrating humans and machines with a goal of augmenting human cognition, can be dated back to the 1950s [3] [10] [24]. The presented model and system are related to past work as follows. Specifically, the section lists different approaches for information transfer from the brain or peripheral nervous system to a computer.

2.1 Invasive Systems

There has been work done in obtaining signals directly from the brain through implanted systems. The work in this domain includes using intracranial electrocorticography (ECoG) to measure neural activity and correlating signals with characters [7] [37]. These systems are used with epilepsy patients who have undergone surgery and undergo temporary placement of electrode arrays on the surface of brain tissue. In addition, there have been efforts to measure local field potentials (LFP) using intracortical electrodes [35] [26]. While these approaches are valuable to rehabilitate totally locked-in patients, the methods suffer from low information transfer rates and particularly slow speeds of communication. These systems have a high degree of invasiveness for these approaches to scale in any practical way beyond a clinical setting, and require a procedure to be performed on the human-user to implant a measuring electrode on the surface of or inside brain tissue. There have been ex-
plorations surrounding measurement of movement of internal speech articulators by placing sensors inside these articulators. Hofe et al. 2013 [19] and Fagan et al. 2008 [11] used permanent magnet (PMA) sensors to capture movement of specific points on muscles used in speech articulation. The approach requires permanent fixing of magnetic beads invasively, and capture movement information when a user mouths word phrases. Sahni et al. 2014 [30] use a combination of magnets to measure tongue motion and an in-ear sensor to measure ear deformation during jaw motions.

In contrast, the system proposed herein measures endogenous signals non-invasively, without discernible movement and with high transfer rates.

2.2 Non-Invasive Systems

There have been multiple approaches proposed to interface humans and machines in a non-invasive manner. Florescu et al. 2010 [13] propose characterization of the vocal tract using ultrasound to achieve silent speech communication. The system achieves good results when combined with a video camera looking directly at the mouth of the user. The obtrusiveness of the approach impedes the scalability of these solutions in real-world settings. Wand et al. 2016 [38] used deep learning on video without acoustic vocalization. The approach requires externally placed cameras to decode language from movement of the lips.

Wand and Schultz 2011 have demonstrated silent speech using a phoneme based acoustic model using surface electromyography (EMG) [39] that measure myoneural activity of the muscles in contact with electrodes (facial muscles). While these methods have high communication speeds, the approach requires overt movements of facial muscles to mouth words, and measures activity at the surface of the skin. Jorgensen et al [22] developed a near-silent system to measure sub-vocal or sub-audible signal that is sotte-voce [21] in characterization of speech, without significant external muscle movement and with relatively little acoustic input. This was done by measuring laryngeal activity during sub-audible speech, and surface signals generated by tongue contact. However, the system is not completely silent at all times, is not real-time and
is unable to recognize alveolar consonants with high accuracy. As an interface, the AlterEgo system differs from the above-mentioned approaches in that it neither requires external muscle movements nor an attempt on the part of the user to produce sound and measures endogenous signals produced internally which are conducted through tissue to the surface of the skin during supralaryngeal articulator activation.

Recently, work has been done in measuring steady state visually evoked potentials (SSVEP), through electroencephalographic (EEG) which measures endogenous brain responses to visual stimulation, when a human-user focuses attention on a visual stimulus at specific frequencies displayed on a screen \[9\] \[42\] \[41\]. While recent SSVEP systems have improved information transfer speeds, these are far below what is required to converse in natural language. These systems enable communication through the user transmitting a single character at a time. Moreover, these systems require an ever-present screen to generate a visual stimulus and an elaborate setup which is difficult to translate to everyday use beyond a clinical setting.

In contrast, the AlterEgo system has a significantly higher information transfer rate than previous listed modalities. An in-depth analysis is presented in Chapter 5, with a comparison of the presented system against a range of modalities.
Chapter 3

Methods and Implementation

3.1 Neural Control of Speech

Speech production is one of the most complex behaviors performed by human beings. Human beings communicate with tremendously high speeds, even during normal conversation. Hence, the production of speech involves a complex series of intricate and coordinated neural events with a high degree of neuro-muscular innervation. There are more than 100 muscles involved in speech, which are optimized for speed and fine control, in contrast to the limb motor system, which is comparatively less complex and optimized for force.

The neural system for motor control is organized in a hierarchical manner. The hierarchy of the motor system includes the Broca’s area, motor association cortex areas, corticospinal and corticonuclear tracts, spinal cord and brainstem nuclei, basal nuclei, the cerebellum, and indirect activation tracts. A speech command once instantiated, is encoded as an instance mediated by areas in the brain, namely the Broca’s area, and subsequently the supplementary motor area to map into muscular movements for articulation [6] [39]. This cortical control for voluntary articulation is enabled through the ventral sensorimotor cortex which controls the activation and activation rate of a motor unit, via projections from the corticonuclear tract to the face, laryngeal muscles, tongue, pharyngeal muscles and other articulators.

Cranial nerves (CN) comprise a part of the peripheral nervous system and emerge
directly from the brain (Figure 3-1). The nerves innervate the oral, pharyngeal and laryngeal musculature among other regions and provide critical motor and sensory information. These consist of efferent (neural pathways that relay commands from central nervous system to a muscle or other end organ) motor fibers that emerge from brainstem nuclei and afferent fibers that arise from peripheral ganglia (peripheral nervous system). The motor nuclei of the cranial nerve receives impulses from the cerebral cortex, which is propagated via the corticonuclear tract (Figure 3-1). The impulse through a neuromuscular junction, triggers an action potential in the muscle fiber resulting in the motor action.
### Table 3.1: Motor and Sensory Functions of Cranial Nerves relevant to Speech and Hearing

<table>
<thead>
<tr>
<th>Cranial Nerve</th>
<th>Motor function</th>
<th>Sensory function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trigeminal Nerve (CN V)</td>
<td>Movements related to biting and chewing</td>
<td>Sensory data from palate, teeth and anterior tongue</td>
</tr>
<tr>
<td>Facial Nerve (CN VII)</td>
<td>Facial muscle movements</td>
<td>Sensation to external ear</td>
</tr>
<tr>
<td>Vestibulocochlear Nerve (CN VIII)</td>
<td>-</td>
<td>Equilibrium and hearing</td>
</tr>
<tr>
<td>Glossopharyngeal Nerve (CN IX)</td>
<td>Elevation of pharynx and larynx</td>
<td>Sensation to posterior tongue and upper pharynx</td>
</tr>
<tr>
<td>Vagus Nerve (CN X)</td>
<td>Elevation of palate, phonation</td>
<td>Sensory data from tongue and external ear</td>
</tr>
<tr>
<td>Hypoglossal Nerve (CN XII)</td>
<td>Movement of tongue</td>
<td>-</td>
</tr>
</tbody>
</table>

The primary efferent signaling, during internal articulation, is through CN IX, CN X, CN XII.

### 3.2 Internal Articulation

Acoustic speech has three central mechanisms, namely: speech respiration, phonation and articulation. Internal articulation, in this text, is described as activation of speech articulators, that are supralaryngeal in nature, without phonation and without engaging facial or laryngeal muscles. This is a significantly attenuated form of speech, that does not engage the vocal folds or produce any acoustic output and is indiscernible to an external observer. Internal articulation of speech could be volitional or non-volitional. However, in this text we primarily focus on volitional internal articulation as a form of human-machine communication.

The full range of muscles neurologically activated in the oral and pharyngeal musculature during internal articulation are as follows: geniohyoid, mylohyoid, genioglossus, superior longitudinal, inferior longitudinal, transverse, vertical, hyoglossus, palatoglossus, styloglossus, levator palatini, musculus uvulae, tensor palatini, palatopharyngeus, superior pharyngeal constrictor, medial pharyngeal constrictor, and inferior pharyngeal constrictor. These muscles comprise the internal speech articulators such as the velum, uvula, tongue and soft palate. The subjective experience
Figure 3-2: **Volumetric Conduction of Signal.** The figure shows a simplified model of an efferent neural signal being conducted through tissue and measured at the electrode. In the model the voltage attenuates as \( V \propto \frac{1}{r} \) which gives an intuition of voltage signal attenuation when measuring electrode and signal source are not in close proximity.

of a human subject during internal articulation, is the subject "talking to herself" while also engaging one or more of these articulators in a highly attenuated manner and the experience of unobservable but volitional form of speech. In contrast to regular overt speech, a human subject’s respiration pattern remains unchanged in most cases during internal articulation.

This ionic movement, caused by muscle fiber resistance during internal muscle activation, generates time-varying potential difference patterns that occur in internal articulators during intention to internally speak, leading to a corresponding myoelectric signature that is recorded by the system described herein, as electromyography (EMG), from the surface of the skin in the absence of acoustic vocalization and facial muscle articulation for speech. The signal (voltage) of efferent activation of a motor unit sourced deep within the oral musculature, is significantly attenuated when it is conducted through bone and tissue, and reaches the surface of the skin (Figure 3-2). This is in contrast to measuring electromyography from the surface of the skin during limb motor control, where the activated limb muscle/signal source is closer to the measuring electrode, which then receives a strong signal.
Figure 3-3: Electrode Target Positions and Associated Muscle Regions. It is important to note that the facial muscles are themselves neurally inactive, and are shown here to clarify positioning of electrodes. These positions were used in our human subject experiments.

The positions of electrodes to measure internal muscle activation were decided empirically in our setup, during pilot studies conducted with 3 human subjects (1 female, $\mu = 29.33$ years). A total of 7 positions were selected as shown in Figure 3-4. Symmetrical equivalents of target locations across the craniocaudal axis were ignored in order to avoid feature repetition, while having an electrode distribution across the lower face and neck. The positioning of measurement electrodes can still be flexible, due to cylindrical symmetry of volumetric conduction of a motor unit signal and the measuring electrodes not being in extreme close proximity of the signal source.

### 3.3 Signal Conditioning and Processing

Signals are captured using electrodes from the target areas shown in Figure 3-4. The system uses TPE plastic, gold plated silver electrodes (1.45 mm diameter conductive area), in combination with Ten20 (polyoxyethylene (20) cetyl ether) conductive paste (Weaver and Company). A reference electrode is placed on the back of the human subject’s earlobe. The arrangement could also be such that the reference is bifurcated into two reference points measuring from back of both earlobes and averaging the reference measurement. An electrode functioning as ground, which inserts interference waveforms to cancel out noise, is placed on back of the other free earlobe. This is
Figure 3-4: Electrode Target Positions in Second Iteration of System. We designed a second iteration of the wearable that would be more feasible to use in quotidian settings. The figure shows the electrode positions in the wearable.

for canceling $\sim 60$ Hz line interferences and to achieve higher signal-to-noise (SNR) ratio. The signals are sampled at 250 Hz and differentially amplified at 24 gain (Texas Instruments, OpenBCI).

A real time signal is mean normalized, taking a specific temporal frame of the signal (the window length in time is a hyperparameter and is subject to calibration, 4 seconds in history in the current system). The signals are fourth order IIR butterworth filtered (1.3 Hz to 10 Hz). The high pass filter is used in order to prevent signal aliasing artifacts. The low pass filter is applied to avoid movement artifacts in the signal. The root mean square of the voltage amplitudes, within the window frame, is computed and then compared against a cutoff voltage to detect internal articulation in real-time. A root mean squared voltage greater than the cutoff would be associated with motor behaviors other than internal articulation (such as facial movements). The cutoff voltage is subject to calibration, and was empirically set (15 $\mu$V in current system). If $V_{rms} < V_{cutoff}$, then the signal is included for further analysis, otherwise the frame is rejected.
3.4 Indiscernible Silent Speech Recognition

We use a modular approach to detect and recognize silent speech from internal articulation in real-time. This is primarily done through two convolutional neural networks. The first network performs activity detection, and deals with extracting relevant windows out of real time noisy data. The second network classifies the incoming detected signals into word labels. An alternative approach could be to implement an end-to-end single network that performs both functions. The current approach was used on account of a simplified implementation, separate evaluation and limited data.

The signal is passed through a spatiotemporal convolutional network (Figure 3-5, and then subsequently followed by a convolutional layer. The non-linearity function is a Rectified Linear Unit (ReLU) at the end of each layer. This is then followed with a batch normalization, and then average pooled. As shown in Figure 3-5, this is then followed by two convolutional layers, a flattened layer and three subsequent fully connected layers. The kernel size of each convolutional layer is 3 with stride 1, the number of kernels are 300. The network is L2 regularized. This network was evaluated with the pilot study human subjects, and achieved an activity detection of 91.46% on a digit vocabulary (words "zero" to "nine") on a frame-by-frame basis binary classification of activity/no activity.

This signal flagged as internal articulation is passed through a 1-dimensional con-
convolutional neural network to classify into word labels with the architecture described as follows (see Figure 3-7). The hidden layer convolves 400 filters of kernel size 3 with stride 1 with the processed input and is then passed through a rectifier nonlinearity. This is subsequently followed by a max pooling layer. This unit is repeated twice before globally max pooling over its input. This is followed by a fully connected layer of dimension 200 passed through a rectifier nonlinearity which is followed by another fully connected layer with a sigmoid activation. The network was regularized using a 50% dropout in each hidden layer to enable the network to generalize better on unseen data. The error during training was evaluated using a cross entropy loss. The neural networks were trained on a single NVIDIA GeForce Titan X GPU. The evaluation of recognition is further expanded through a human subject experiments, discussed in Chapter 5.

Figure 3-6: Architecture of Convolutional Neural Network classifying Internally Articulated Speech into Word Labels.
Figure 3-7: Sensorimotor Feedback during Speech Articulation. This shows the schematic of the Directions into Velocities of Articulators model of speech acquisition and production (from Guenther et al. 2006, Cortical interactions underlying the production of speech sounds [16]). In the silent speech interface, speech is internally articulated without phonation and sensory feedback is given through bone conduction.

Figure 3-8: Hardware Arrangement of the AlterEgo Wearable.
3.5 Closed-Loop Interface

Silent speech recognition in the AlterEgo system attempts to open up a unique opportunity to enable real-time bi-directional human-machine communication in a concealed, seamless manner, where the element of interaction is in natural language. The motor action of speech articulators during speech, provides sensorimotor feedback to plan and correct further motor control and thereby the next word communicated [20] [18]. This is experienced concomitant with inner speech. In the AlterEgo setup, this is subjectively experienced through volitional internal articulation and aural feedback through bone-conduction audio. We use bone conduction transducers as the aural output, so as to not impede the user's sense of hearing and give an aural feedback of the internally articulated word recognized by the system.

An important aspect of such an interface is that it is both non-invasive and non-intrusive to user privacy. A primary focus of the system design is not to render ethics as an afterthought. Volitional internally articulated speech enables private human-machine communication, but also provides a way to do so without a machine system having continuous access to brain activity, and a volitional transfer of information. An interface such as this potentially facilitates a complementary synergy between human users and machines, where certain tasks could be outsourced to a computer while the computation still seeming as "intrinsic" to the human user. After an internally articulated word is recognized, the computer contextually processes the phrase according to the relevant application the user accesses (e.g. - An IoT application would assign the internally vocalized digit 3 to device number 3 whereas the Mathematics application would consider the same input as the actual number 3). The output, thus computed by the application, is then converted using Text-to-Speech and aurally transmitted to the user. In the final iteration of the system (Figure 3-10), bone conduction transducers are embedded in the portion of the wearable behind the ear, and transmit audio information to the inner cochlea without disruption normal hearing. The computation happens online, through an internet connected server that communicates with the device both to receive input and transmit output to the bone
conduction transducers (Figure 3-8).
Figure 3-10: Second Iteration of the AlterEgo Prototype.
Figure 3-11: Flowchart of the Silent Speech Interface. The figure shows the various steps of the system. The step when various features are calculated for feature fusion is an alternate strategy, which feeds computed features into the network.
Chapter 4

Evaluation

This chapter describes the results of our experimental evaluation. The experiments presented were designed to demonstrate the system’s potential to enable volitional and internal human-machine communication with significant robustness and communication speeds that facilitate a real-time human-machine duplex interface.

4.1 Silent Speech Recognition

We first sought to evaluate the word accuracy (WAcc) of the system and the recognition model, through a multi-user experimental study. The experiments were performed with informed consent at the Media Lab, Massachusetts Institute of Technology, and the approval of the Committee on the Use of Humans as Experimental Subjects. The experiments were conducted with 10 recruited participants (6 female) between 19 and 31 years old ($\mu = 23.6$ years). The recruited participants had no history of a neurological or speech-related pathology. The participants had no prior experience with the system evaluated in the study.

The electrodes were placed and adjusted acutely on the surface of a participant’s skin, via the wearable, according to electrode placement positions described in Chapter 3. The experimental trials were conducted using the same electrode placement protocol for all participants. The electrodes were made of a coating of conductive paste (Ten20, Weaver and Company). The experiments were conducted in a station-
ary setting, with the participant seated in front of a display monitor, meant to display instructive text.

The participant was first made accustomed to internally articulated speech. This was done during a brief initiation period, where the subject would read a passage displayed on the screen. The subject would be first asked to read the text using overt vocalized speech. The subject would then attempt to attenuate speech, initially mouthing words without overt phonation and then further reducing movement until there were no discernible movements. The participants were instructed to internally articulate words, without encoding the words into coded mannerisms of articulation, and in a natural conversational manner. The subject would sometimes be given feedback on the screen to indicate neurological activation of internal articulatory muscles to help prepare the subject. In some cases, the subject was also given mirrored video feedback to adjust her speech, when attenuating their speech, if the subject was moving or not.

The arithmetic computation application was used as the vocabulary for accuracy evaluation. For each participant, a total of 750 digits were displayed, randomly sequenced on a computer screen. The digits were randomly chosen from a total of 10 digits (0 to 9), such that each digit exactly appeared 75 times. Each participant had a push-button switch, sending high-voltage spikes to mark beginning and end of internally articulated word.

The experiment trial data for each participant was split according to an 80/20 random split for training and testing for each user. The data was used to train models as described in Chapter 3. The word accuracy (WAcc) values for each user were recorded for 10 runs of leave-group-out cross-validation. Figure shows the word accuracy distribution. The average WAcc across all experiments participants and leave-group-out cross-validation is 92.01%. We conducted live testing for each user to observe the latency of the system for real-time predictions. The latency refers to the computational latency of the silent speech system as measured from the end of an utterance until the corresponding transcription is produced. The average latency for the 10 users was 0.427 seconds (3sf).
Figure 4-1: **Word Accuracy of Silent Speech Recognition.** The model is evaluated on 10 users, using 10 runs of cross-validation of the dataset with arithmetic digits (0-9). The edges of boxes represent 1st and 3rd quartile respectively, and the whiskers extend to 100% coverage.
4.2 Information Rate

We are interested in evaluating whether the AlterEgo system could enable a real-time human-computer communication channel. The information transfer rate (ITR) was calculated (in bits per second) for the variant of the system used in the aforementioned experimental setup, as follows [27]:

\[
ITR = \left( \log_2 N + P \log_2 P + (1 - P) \log_2 \left[ \frac{1 - P}{N - 1} \right] \right) / T
\] (4.1)

where \(N\) is the number of possible states or targets, \(P\) is the classification accuracy, and \(T\) is the time (in seconds per symbol) to select a state or a target. The system evaluated was memoryless, i.e. each state or word is assumed independent. The marginal probabilities were uniform over the state space.

The participants were given a visual cue, to internally articulate the word indicated. The online system would gauge the trial duration, through a push-button indication by the participant during internal articulation, as shown in Fig 4-2. We used the same targets (0-9 digits), as used in the previous experiment. A few practice sessions were conducted with the participants (~ 15-30 min.) with the online system, after the system training period. In sum, we conducted 10 trials for each participant, and used the mean trial duration \(T\) for computing ITR of each participant. The results of the experiment are shown in Table 4.1. The average ITR in the session was 10.302 ± 1.47 bps, across all participants. The maximal and minimal ITR was 11.913 bps (P2) and 7.639 bps (P7), respectively. Table 4.2 lists the ITR of brain-computer interface systems and modalities. The ITR values of previous systems are representative of the values reported for each category [9] [31] [29] [8] [36] [12] [40] [14] [25] [15] [7] [28] [4].

Consider the reported ITR values in the table. While there has been a clear trend in improvement of ITR, the table shows that there is a marked difference when comparing the average ITR of the current system, against the ITR of a range of previous modalities listed (10.302 bps vs. 4.50 bps, Table 4.2). To our knowledge, the current ITR represents the highest reported in invasive or non-invasive brain- and
Figure 4-2: **Online System measuring Information Transfer Rate.** A visual cue of a word is displayed on screen. The rising edge of voltage coincides with the start of the subject internally articulating the word through a push-down button switch and held through the duration of articulation. The falling edge coincides with the end of articulation. The raw data during the period of high signal ($T$) is then further processed for recognition.

Peripheral nerve-computer interfaces.

A significant drawback of brain- and peripheral nerve-computer interfaces has been low speed of communication. The average time of word communication in the current system is sub-second. The average trial duration values, in Table 4.1 for each participant, are in the range of natural language conversational speeds [reference], which follows from similar neuromuscular process involved in overt and internally articulated speech. The symbol communication speeds of the current system enable a real-time interface.

We note that the number of distinguishable symbols in the current experimental setup (0-9 digits) is restricted, and could be significantly improved upon. Nonetheless, in the current AlterEgo system, expanding the number of symbols or vocabulary size by 2 (from 10 to 12 symbols), while assuming similar system communication speeds and robustness, would increase the ITR by $\sim 1$ bps:

---

1In a practical and full conversational scenario, we expect the rate of internal articulation to exceed that of overt speech. One reason for the possible difference, is that internally articulated speech would not be interrupted due to speech respiration stops and pulmonic egressive air flows.
<table>
<thead>
<tr>
<th>Participant</th>
<th>Trial Duration, ms</th>
<th>ITR, bps</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>235</td>
<td>11.345</td>
</tr>
<tr>
<td>P2</td>
<td>223.8</td>
<td>11.913</td>
</tr>
<tr>
<td>P3</td>
<td>295.7</td>
<td>9.016</td>
</tr>
<tr>
<td>P4</td>
<td>295.5</td>
<td>9.022</td>
</tr>
<tr>
<td>P5</td>
<td>228.6</td>
<td>11.663</td>
</tr>
<tr>
<td>P6</td>
<td>277.2</td>
<td>9.619</td>
</tr>
<tr>
<td>P7</td>
<td>349</td>
<td>7.639</td>
</tr>
<tr>
<td>P8</td>
<td>227.2</td>
<td>11.735</td>
</tr>
<tr>
<td>P9</td>
<td>257</td>
<td>10.374</td>
</tr>
<tr>
<td>P10</td>
<td>227</td>
<td>11.745</td>
</tr>
<tr>
<td>Mean ± SD</td>
<td>-</td>
<td>10.302 ± 1.47</td>
</tr>
</tbody>
</table>

Table 4.1: **Information Transfer Rate of AlterEgo.** The table shows the average ITR (in bps) of the system, with detection of 10 words ($N = 10$). An average ITR was computed for each participant over 10 experiment trials for each participant. The table shows average trial duration (in ms) for each participant.

<table>
<thead>
<tr>
<th>Platform</th>
<th>Category</th>
<th>ITR, bps</th>
</tr>
</thead>
<tbody>
<tr>
<td>AlterEgo</td>
<td>Non-Invasive</td>
<td>10.302</td>
</tr>
<tr>
<td>joint frequency-phase modulated SSVEP</td>
<td>Non-Invasive</td>
<td>4.50</td>
</tr>
<tr>
<td>multi-neuron neural prosthesis</td>
<td>Invasive</td>
<td>3.50</td>
</tr>
<tr>
<td>code modulated VEP</td>
<td>Non-Invasive</td>
<td>1.91</td>
</tr>
<tr>
<td>electrocorticogram P300</td>
<td>Invasive</td>
<td>1.90</td>
</tr>
<tr>
<td>SSVEP</td>
<td>Non-Invasive</td>
<td>1.44</td>
</tr>
<tr>
<td>Hybrid SSVEP-EMG</td>
<td>Non-Invasive</td>
<td>1.39</td>
</tr>
<tr>
<td>fMRI</td>
<td>Non-Invasive</td>
<td>0.60</td>
</tr>
<tr>
<td>Forearm EMG</td>
<td>Non-Invasive</td>
<td>0.51</td>
</tr>
<tr>
<td>P300</td>
<td>Non-Invasive</td>
<td>0.50</td>
</tr>
<tr>
<td>Hybrid EEG-EOG</td>
<td>Non-Invasive</td>
<td>0.41</td>
</tr>
<tr>
<td>EOG</td>
<td>Non-Invasive</td>
<td>0.32</td>
</tr>
<tr>
<td>fNIRS</td>
<td>Non-Invasive</td>
<td>0.06</td>
</tr>
</tbody>
</table>

Table 4.2: **Information Transfer Rates of Brain- and Peripheral Nerve-Computer Interfaces.** The table lists previously reported ITR across a range of systems.
\[ \Delta ITR \approx \frac{NP - 1}{N(N - 1) \ln 2} \frac{\Delta N}{T} \] (4.2)

where \( \Delta ITR \) and \( \Delta N \) represent change in ITR and change in number of words or symbols, respectively (Appendix). There is an immense potential to further increase vocabulary and ITR multifold. In addition, an aspect not reflected in the ITR is that the symbols in the current setup are words, unlike previous BCI spellers which use individual letters as targets/symbols.
Chapter 5

Application Examples

This chapter presents some application examples using the model and platform described in this thesis. The examples are intended more to show computing as a cognitive extension, through the human-machine duplex interface, than to discuss the particulars of the application area.

It could be shown that a bi-directional interface could assist in and externalize various complex human cognitive processes such as knowledge formation, memory, evaluation, reasoning and computation, comprehension and production of language, and problem solving and decision making. The examples described in this chapter are the ones that have been implemented in some form, using the AlterEgo system, and show versatility of the platform.

5.1 Closed-Loop Computing

This section discusses the set of situations where a human user poses a query or a computation to the machine, and receives resultant information through aural feedback.

The first example in this category pertains to externalizing arithmetic computations through the interface. A human-user would internally articulate an arithmetic problem, communicating with a machine, where the machine would compute a resul-
Coalesced Human-AI Decision Making in the game of Chess.

The figure shows a user internally communicating the previous move played by an opponent, to a game playing AI agent running in the background, and seeks counsel on next move to be played. The AI agent optimizes for the next move and relays the information to the user, through bone-conduction.

In the present implementation, including activity detection, an inactivity of 1 second has been set to indicate end of a word. So a user would internally articulate "three pause add pause five" to get a resultant audio answer of "eight". An inactivity of 3 seconds indicates end of expression.

The second example is a demonstration of how a human-user’s decision making
could be offloaded to machine intelligence, through the interface. To demonstrate this, we use the setup of Chess and Go. Chess and Go are known to have high state-space and game-tree complexity and have exponential run-time, and thus have long been popular decision problem environments for machine intelligence algorithms to tackle, to show sophistication of the methods [33] [34]. Figure 5-1 shows our setup, where the human-user communicates the game state through internal articulation to an artificial intelligent (AI) agent, and receives counsel aurally on next move to be played. This was an example showing how artificial intelligence decision making agents could serve as cognitive adjuncts, where a complex decision problem is posed by the human-user and collaboratively solved through such a bi-directional interface. In the current implementation, the user would internally articulate the previous move played by the opponent, and the agent would communicate the next move to be played.

Language comprehension and production are central human cognitive skills [32]. The following is an example that has not been implemented in full but demonstrates a use-case. In this example, a human-user internally communicates an expression in a known language and then transmits name of a target language. The machine communicates the translation of the phrase back to the user. For instance, a user would internally articulate ”Hello, how are you?” and ”French” and the machine would communicate ”Salut, comment allez-vous?” through bone-conduction transducers or peripheral speakers or directly to another AlterEgo system in real-time. This example is to show how such a duplex interface could enable multi-language comprehension and communication exist seemingly as innate skills to the user, without a requirement of learning the language. It follows then that the system, through such communication, could potentially assist in second-language learning. However, the current implementation of the application in the AlterEgo system is limited in its vocabulary, and translates 10 English phrases in 5 languages. A deployable system would ideally have a word vocabulary close to the span of common conversational language.

A significant increase in symbols/vocabulary recognized by the system would also enable augmentation of the human-user’s memory, where semantic information could be internally articulated and retrieved at a later time.
Figure 5-2: **Controlling IoT Devices through AlterEgo by Internally Articulating Commands.** The figure shows the user (left-right) internally articulating the phrases "bulb 1 2 on". The system recognizes the commands, sending wireless instructions to an Internet enabled switch, that switches on bulb 1 and bulb 2. This example demonstrates how remote control becomes endogenous to the user.

### 5.2 Open-Loop Computing

This section categorizes the cases where human-machine communication is simplex for the most part, and when the human requires no aural feedback.

This example shows how a human-user could be directly connected to Internet-of-Things (IoT) devices, and control these using internally articulated speech. Figure 5-2 shows a human user, using AlterEgo, controlling room lighting using internally articulated speech, without any observable action. The system can be used to control appliances and devices such as HVAC systems, television control etc. The current implementation could be calibrated and trained to access specific services. For instance, a human-user could train the phrase "Cab to home" to book ride-sharing transport from the current location using the interface. Such an interface would also be useful for input in Virtual Reality/Augmented Reality applications.

### 5.3 Telecommunications

A duplex interface would also establish a bi-directional channel for direct communication between people. Human-users would send and receive messages without voice and without any observable action, and simply by internally articulating the message. To realize this in its full capacity, the vocabulary would need to be further expanded
by a considerable amount. In the current implementation of the AlterEgo system a user can articulate 5 common conversational sentences through the interface, as text message to another person.
Chapter 6

Future Work

There remain multiple domains for future work. In this chapter, we identify some of those areas for further work on the system and the interface.

6.1 Vocabulary Recognition

A substantial increase in range of symbols or phrases recognized by the system, with an increased ITR, would enable a further range of applications. To that end, there is room for potential improvement of the AlterEgo system. The iteration of the system described in the thesis, classifies individual words or symbols. A sentence based recognition, with vocabulary resembling the span of conversational language, remains an ambitious goal that would enable full and internal human-machine/human communication.

A successful implementation of an interface with increased vocabulary would require access to larger amounts of example data and an augmentation of the detection and recognition models described in this thesis. Similar to any data driven technique, availability of data remains a bottleneck in creating a robust generalized model. The future implementations of the system would benefit from human experiments and experimental data. This would help in creating a sequence to sequence recognition model that translates efferent neural signals into sentence level transcription.
Moreover, the current system requires individual training for each individual. A future system implementation should include algorithms and models that would transfer across multiple users with high recognition accuracy, to reduce or omit individual user calibration and training time.

6.2 Physiological Rehabilitation and Prosthesis

There is significant potential in using the described system to rehabilitate people with neurological or speech-related pathologies. There have been no formal experiments conducted using the described system, with human subjects with such conditions, yet. This presents an opportunity to investigate the use of such a communication channel to patients with a form of neurological conditions such as amyotrophic lateral sclerosis (ALS), Parkinson's disease, stroke. The advantages of using such a system over previous systems, for instance steady-state visually evoked potentials (SSVEP) based systems, is that the current system does not require volitional eye gaze and an ever-present feedback screen. The current system has a higher ITR, uses natural conversation and has a high feasibility of being deployed in an ambulatory and non-clinical setting, in comparison to previous systems. Just as in the case with SSVEP based systems, totally locked-in patients would not be able to use the current system. The user would need to have some volitional motor control of speech articulators. We imagine that a future version of the interface, with both input and auditory feedback, could be valuable to people with speech conditions such as stuttering, cluttering, dysarthria using speech synthesis and assistive feedback. The interface could provide useful as an assistive system for Alzheimer's disease patients in certain scenarios, as an internal memory prosthetic through recording and retrieval of semantic information, through internally articulated speech. The system could potentially be used as a speech prosthetic for laryngeal cancer patients who have had their voice-box removed, and for speech recovery in post-operation laryngectomy.
6.3 Ambulatory Conditions and Testing

We hope to further develop the current AlterEgo system for deployment as a duplex interface in real-world situations. In such scenarios, the system would be need to accurately distinguish between internal articulated speech and other neuromuscular signatures in general motor control. The system would likely be developed to distinguish between volitional and non-volitional internally articulated speech. While the current system has been instantiated as a wireless wearable, the current iteration has not been optimized and evaluated for mobile conditions. The system would need to be tested through longitudinal tests in human subjects for usability in quotidian conditions.

6.4 Suprasegmental Elements

An area of future work would be to investigate if suprasegmental elements of speech, such as intonation and stress could be reconstructed or predicted directly from internal articulatory speech signals. Direct mapping of raw efferent neural signals to speech would have implications for telecommunications and in rehabilitating patients bereft of voice.

6.5 User Interface Design

There is much future work to be done in the area of designing a seamless interface. There are multiple design aspects that would need to be considered. The timing of the output of the interface should not be disruptive. Another aspect that would need to be optimized is the situation when the user would need to make corrections during an instance of incorrect word prediction. The audio feedback during word recognition might be useful only with minimal latency and be disruptive otherwise. These situations would need to be designed for and evaluated through user studies.
Appendix A

A.1 Change in Information Transfer Rate

The change in Information Transfer Rate (ITR), when keeping parameters constant, could be simply computed by plugging in all parameter values. Nevertheless, a change in ITR as a function of change in parameter values could be given by:

\[
df(x_1, x_2, x_3, \ldots, x_n) = \frac{\partial f}{\partial x_1} dx_1 + \frac{\partial f}{\partial x_2} dx_2 + \frac{\partial f}{\partial x_3} dx_3 + \ldots + \frac{\partial f}{\partial x_n} dx_n
\]

ITR is given as:

\[
ITR = \left( \log_2 N + P \log_2 P + (1 - P) \log_2 \left[ \frac{1 - P}{N - 1} \right] \right) / T
\]

\[
dITR(N, P, T) = \frac{\partial ITR}{\partial N} dN + \frac{\partial ITR}{\partial P} dP + \frac{\partial ITR}{\partial T} dT
\]

Assuming change in only N:

\[
\Delta ITR \approx \frac{NP - 1}{N(N - 1) \ln 2} \frac{\Delta N}{T}
\]
Bibliography


