AlterEgo: A Personalized Wearable Silent Speech Interface

Arnav Kapur
MIT Media Lab
Cambridge, USA
arnavk@media.mit.edu

Shreyas Kapur
MIT Media Lab
Cambridge, USA
shreyask@mit.edu

Pattie Maes
MIT Media Lab
Cambridge, USA
pattie@media.mit.edu

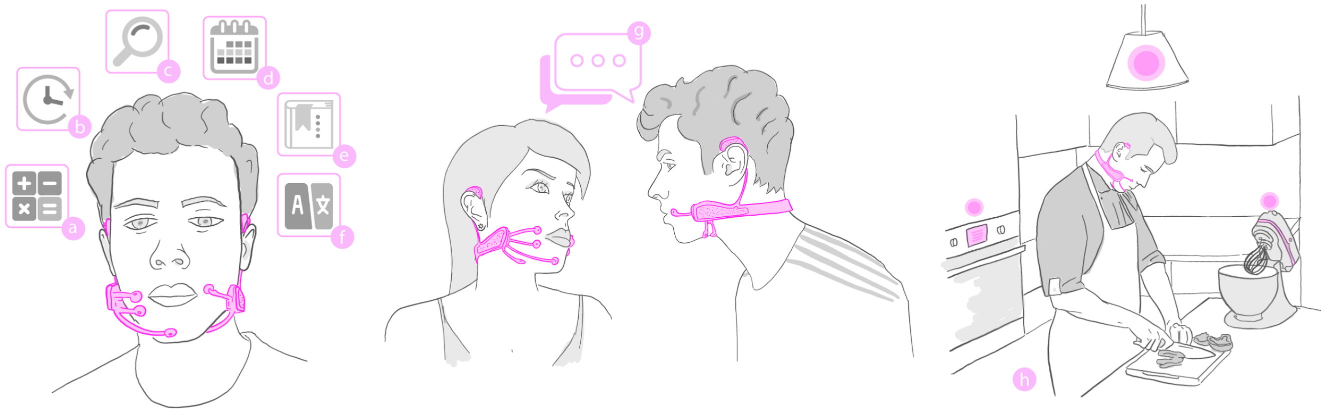

Figure 1. Alterego seeks to make computing a natural extension of the user's own cognition by enabling a *silent*, *discreet* and *seamless* conversation with machines and people, in likeness to the user talking to her own self.

ABSTRACT

We present a wearable interface 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 intention to speak and internal speech is characterized by neuromuscular signals in internal speech articulators that are captured by the AlterEgo system to reconstruct this speech. We use this to facilitate a natural language user interface, where users can silently communicate in natural language and receive aural output (e.g - bone conduction headphones), thereby enabling a discreet, bi-directional interface with a computing device, and providing a seamless form of intelligence augmentation. The paper describes the architecture, design, implementation and operation of the entire system. We demonstrate robustness of the system through user studies and report 92% median word accuracy levels.

Author Keywords

Silent Speech Interface; Intelligence Augmentation; Peripheral Nerve Interface; Human-Machine Symbiosis

IUI'18, March 7–11, 2018, Tokyo, Japan

ACM ISBN 978-1-4503-4945-1/18/03$15.00
https://doi.org/10.1145/3172944.3172977

INTRODUCTION

The vision of closely coupling humans and machines has been advanced and re-imagined in successive iterations. Input devices have come a long way since punchcards and present day input modalities have enabled computing devices to become an intrinsic parts of our lives. Keyboards (or typewriter style input devices) 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 inputs 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 [5,13,21–23,27]. Despite significant advances made, machines, input modalities and interactivity still exist as *external* artifacts to the human user, in an obstacle to fully realize symbiosis of humans and machines.

We present AlterEgo a wearable silent speech interface that allows a user to provide arbitrary text input to a computing device or other people using natural language, without discernible muscle movements and without any voice. This allows the user to communicate to their computing devices in natural language without any observable action at all and without explicitly saying anything.

In summary, this paper makes three primary contributions:

1. We introduce a novel wearable architecture for a bi-directional silent speech device.

2. We outline the neuromuscular input needed for detecting silent speech.

3. We demonstrate the feasibility of such silent speech recognition based on neural information, and demonstrate the utility of the device as a personal computing platform and interface.

AlterEgo, allows people to privately and seamlessly communicate with their personal computing devices, services and other people, such that users leverage the power of computing in their daily lives, as a natural adjunct to their own cognitive abilities, without "replacing" or "obstructing" these abilities.

BACKGROUND AND RELATED WORK

Voice Interfaces

Conversational interfaces currently exist in multiple forms. The recent advances in speech recognition methods have enabled users to have interaction with a computing device in natural language [1,12]. This has facilitated the advent of ubiquitous natural voice interfaces, currently deployed in mobile computational devices as virtual assistants (e.g- Siri [28], Alexa [29], Cortana [30] etc.). These interfaces have also been embedded in other devices such as smart-wearables, dedicated hardware speakers (e.g - Google Home [31], Amazon Echo[32]), and social robots. Another broad category under voice interfaces are modern telecommunications devices for person-person communication (e.g - smartphones, Skype etc). Although, all the aforementioned platforms offer robust voice based interaction, they share common limitations. There are fundamental impediments to current speech interfaces that limit the possibility of their adoption as a primary human-machine interface. We list a few here amongst others:

Privacy of conversation: Speech is broadcasted to the environment by the user when communicating via these interfaces and therefore user privacy is not maintained (e.g - a phone call with another person; communicating with Siri etc.).

Eavesdropping: Voice interfaces are always listening in on conversations, when not desired, only to be visibly activated later on by a specific trigger-word (e.g - 'Ok Google' activates Google Assistant but the application is on nevertheless).

Impersonal devices: These devices are not personal devices and any other user can intentionally or unintentionally send valid voice inputs to these devices.

Attention requiring: Current voice interaction devices have low usability as a device, a user cannot use a speech interface hands free on-the-go, which is the case oftentimes with immobile dedicated speech devices (e.g - Amazon Echo). Moreover, user proximity to the device is required for optimal speech recognition (telecommunications devices).

Silent Speech Interfaces

There have been several previous attempts at achieving silent speech communication. These systems can be categorized under two primary approaches: invasive and non-invasive systems.

Invasive Systems

Brumberg et al. 2010 [6] used direct brain implants in the speech motor cortex to achieve silent speech recognition, demonstrating reasonable accuracies on limited vocabulary datasets. There have been explorations surrounding measurement of movement of internal speech articulators by placing sensors inside these articulators. Hueber et al. 2008 [17] used sensors placed on the tongue to measure tongue movements. Hofe et al. 2013 [16] and Fagan et al. 2008 [9] 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 which does not scale well in a real-world setting. Florescu et al. 2010 [10] propose characterization of the vocal tract using ultrasound to achieve silent speech. The system only achieves good results when combined with a video camera looking directly at the user's mouth. The invasiveness and obtrusiveness or the immobility of the apparatus impedes the scalability of these solutions in real-world settings, beyond clinical scenarios.

Non-Invasive Systems

There have been multiple approaches proposed to detect and recognize silent speech in a non-invasive manner. Porbadnik et al. 2009 [24] used EEG sensors for silent speech recognition, but suffered from low signal-to-noise ratio to robustly detect speech formation and thereby encountered poor performance. Wand et al. 2016 [26] used deep learning on video without acoustic vocalization but requires externally placed cameras to decode language from movement of the lips. Hirahara et al. [15] use Non-Audible Murmur microphone to digitally transform signals. There have been instances of decoding speech from facial muscles movements using surface electromyography.

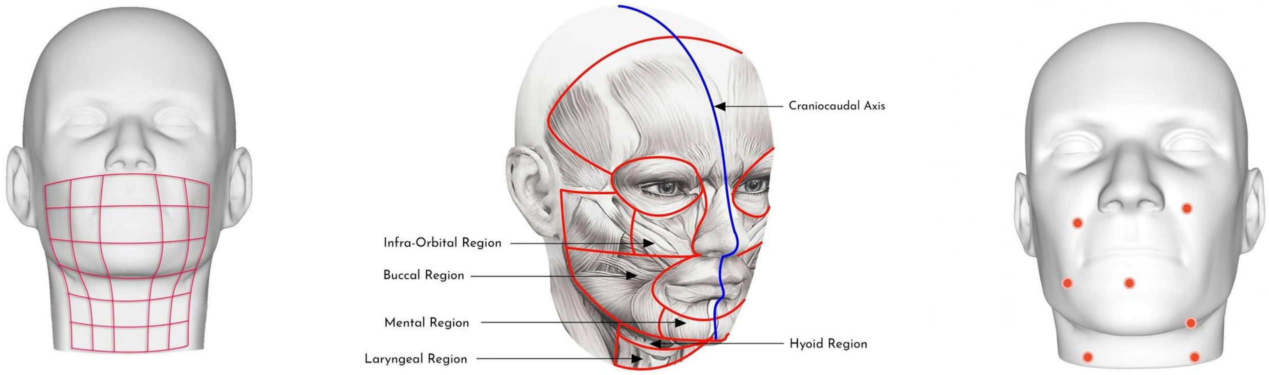

Figure 2. Selection of the final electrode target areas (right) through feature selection on muscular areas of interest (left, center).

Wand and Schultz 2011[25] have demonstrated surface EMG silent speech using a phoneme based acoustic model, but the user has to explicitly mouth the words and have to use pronounced facial movements. Jorgensen et al.[18] use surface EMG to detect subvocal words with accuracy fluctuating down to 33% with the system also unable to recognize alveolar consonants with high accuracy, which is a significant obstruction to actual usage as a speech interface.

ALTEREGO

AlterEgo is a wearable silent speech interface that enables a discreet, seamless and bi-directional communication with a computing device in natural language without discernible movements or voice input (Figure 3-4).

We distinguish our system based on the following points:

1. The first difference is the non-invasiveness of the approach described herein. The system captures neuromuscular signals from the surface of the user's skin via a wearable mask.

2. The existing non-invasive real-time methods with robust accuracies require the user to explicitly mouth their speech with pronounced, apparent facial movements. The key difference between our system and existing approaches is that our system performs robustly even when the user does not open their mouth, make any sound and without the need for any deliberate and coded muscle articulation that is often used when using surface EMG to detect silent speech. The modality of natural language communication without any discernible movement is key, since it allows for a seamless and discreet interface.

3. On a standard digit recognition test, our system achieves a median accuracy of 92%, outperforming conventional methods mentioned above amongst others. Moreover, this is despite the AlterEgo system not requiring any facial muscle movement as opposed to conventional methods that require the user to lip sync the words in a pronounced fashion.

4. This leads to the fourth key point, which is the portability of the wearable device. The proposed device is an ambulatory wearable system which a user just needs to wear for it to function and the device connects wirelessly over Bluetooth to any external computing device.

5. Unlike proposed traditional brain computer interfaces (BCI), such as head based EEG/fMRI/DOT/fNIRS, the platform does not have access to private information or thoughts and the input, in this case, is voluntary on the user's part. The proposed platform is robust on extended vocabulary sizes than traditional BCI since we propose a peripheral nerve interface by taking measurements from the facial and neck area, which allows for silent speech signals to be distilled without being accompanied by electrical noise from the frontal lobe of the cerebral cortex.

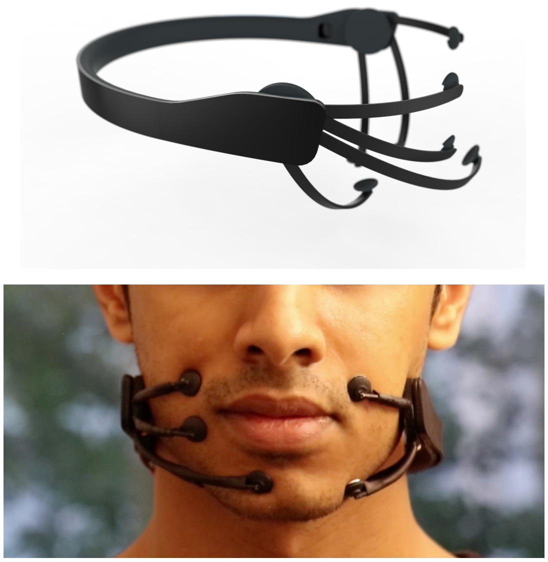

Figure 3. Rendering of the AlterEgo wearable (Top). Front view of the user wearing the device (Bottom).

INDISCERNIBLE SILENT SPEECH RECOGNITION

Internal vocalization, in this text is described as the characteristic inner voice in humans that is usually noticeable while reading and can be voluntarily triggered while speaking to oneself [4], excluding deliberate lip or discernible muscle movement. This is characterized by subtle movements of internal speech articulators.

Speech Synthesis and Electrophysiology

The production of acoustic speech involves a series of intricate and coordinated events and is considered one of the most complex motor actions humans perform. An expression once conceived in the brain, is encoded as a linguistic instance mediated by areas in the brain, namely the Broca's area, and subsequently the supplementary motor area to map into muscular movements for vocal articulation. 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 corticobulbar tract to the face, laryngeal cavity, pharynx and the oral cavity. A motor neuron receives nerve impulses from anterior horn cells in the spinal cord which are propagated to neuromuscular junctions, where a single neuron innervates multiple muscle fibers.

The propagation of a nerve impulse through a neuromuscular junction causes the neurotransmitter acetylcholine to be released into the synapse. Acetylcholine binds with nicotinic receptors leading to ion channels releasing sodium cations in the muscle fiber, triggering an action potential propagation in the muscle fiber. This ionic movement, caused by muscle fiber resistance, generates time-varying potential difference patterns that occur in the facial and neck muscles while intending to speak, leading to a corresponding myoelectric signature that is detected by the system described in this paper, from the surface of the skin in the absence of acoustic vocalization and facial muscle articulation for speech.

Amongst the various muscle articulators involved in speech production [14], we focused our investigation on the laryngeal and hyoid regions along with the buccal, mental, oral and infraorbital regions to detect signal signatures in a non-invasive manner. To determine the spatial locations of detection points, we selected 7 target areas on the skin for detection, from an initial 30-point grid spatially covering the aforementioned select regions. The selection was done on experimental data recorded on which we expand on in following sections. We ranked potential target locations according to the χ^2 filter-based feature ranking, evaluating how signals sourced from each target were able to better differentiate between word labels in our dataset. Symmetrical equivalents of target locations across the craniocaudal axis were ignored in order to avoid feature repetition. In the current iteration of the device, the signals are sourced as 7 channels from the following areas - the laryngeal region, hyoid region, levator anguli oris, orbicularis oris, platysma, anterior belly of the digastric, mentum. The finer positions of the electrodes on the skin, within the selected regions, were then adjusted empirically.

Ranking	Region
1	Mental
2	Inner laryngeal
3	Outer laryngeal
4	Hyoid
5	Inner infra-orbital
6	Outer infra-orbital
7	Buccal

Table 1. Top muscle regions ranked according to the χ^2 filter ranking against a binary-labelled dataset, evaluated in the pilot user study.

Signal Capture, Processing and Hardware

Signals are captured using electrodes from the above-mentioned target areas. In two versions of the device, the device houses either 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) for reduced contact impedance, or passive dry Ag/AgCl electrodes (4 mm diameter conductive area). Although both electrode forms can be integrated into the system, the former offer superior data quality. Therefore, we report experiments, data collection and results based on the former in this manuscript, as a controlled variable. A reference electrode is placed either on the wrist or the earlobe. We use bias based signal cancellation for canceling ~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).

We integrated an opto-isolated external trigger, acting as a final channel stream with high voltage pulses marking starting and ending events of a silent phrase. Subsequently, the signal streams are wirelessly sent to an external computing device for further processing. The signals go through multiple preprocessing stages. The signals are fourth order IIR butterworth filtered (1.3 Hz to 50 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. A notch filter is applied at 60 Hz to nullify line interference in hardware. The notch filter is applied, despite the butterworth filter, because of the gentle roll-off attenuation of the latter.

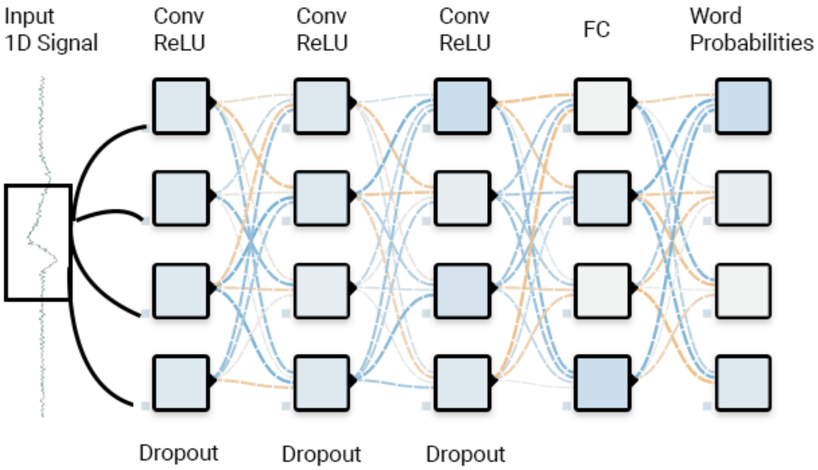

Figure 4. Architecture of the AlterEgo silent speech recognition model.

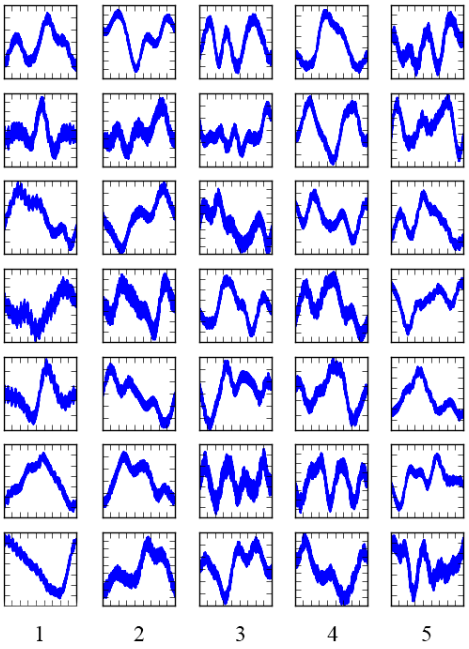

Figure 5. Output of activation maximization on the classes *1-5*, to visualize input signals that maximize the corresponding output activations.

We separate the signal streams into components through Independent Component Analysis (ICA) to further remove movement artifacts. The signals are digitally rectified, normalized to a range of 0 to 1 and concatenated as integer streams. The streams are sent to a mobile computational device through Bluetooth LE, which subsequently sends the data to the server hosting the recognition model to classify silent words. This protocol is in consideration of data transfer speed requirements and power requirements for the system to potentially scale as a wearable device.

Data Collection and Corpus

We created a data corpus comprising datasets of varied vocabulary sizes. Data was collected during two main phases. First, we conducted a pilot study with 3 participants (1 female, average age of 29.33 years) to investigate feasibility of signal detection and to determine electrode positioning. The preliminary dataset recorded with the participants was binary, with the world labels being *yes* and *no*. The vocabulary set was gradually augmented to accommodate more words - for instance, the phonetically dissimilar words *reply*, *call*, *you* and *later* formed another dataset. In sum, the data collected during the study has ~5 hours of internally vocalized text.

In the second phase, we set out to create a data corpus to be used to train a classifier (same 3 participants). The corpus has ~31 hours of silently spoken text recorded in different sessions to be able to regularize the recognition model for session independence. The corpus comprises of multiple datasets (Table 2). In one category, the word labels are numerical digits (0-9) along with fundamental mathematical operations (*times*, *divide*, *add*, *subtract* and *percent*) to facilitate externalizing arithmetic computations through the interface. We expand on other dataset categories in later sections. We use the external trigger signal to slice the data into word instances. In each recording session, signals were recorded for randomly chosen words from a specific vocabulary set. This data is used to train the recognition model for various applications, on which we expand on in following sections.

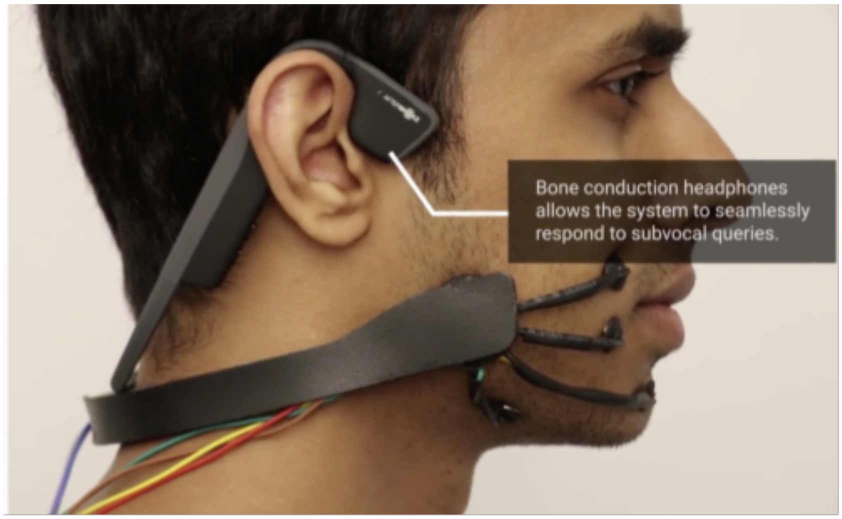
Bone conduction headphones allows the system to seamlessly respond to subvocal queries.

Figure 6. Bone conduction aural output of the AlterEgo system, making it a closed-loop input-output platform.

Silent Speech Recognition Model

The signal undergoes a representation transformation before being input to the recognition model. We use a running window average to identify and omit single spikes (> 30 μVabove baseline) in the stream, with amplitudes greater than average values for nearest 4 points before and after. We use mel-frequency cepstral coefficient based representations to closely characterize the envelopes of human speech. The signal stream is framed into 0.025s windows, with a 0.01s step between successive windows, followed by a periodogram estimate computation of the power spectrum for each frame. We apply a Discrete Cosine Transform (DCT) to the log of the mel filterbank applied to the power spectra. This allows for us to effectively learn directly from the processed signal without needing to hand-pick any features. This feature representation is passed through a 1-dimensional convolutional neural network to classify into word labels with the architecture described as follows. 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 optimized using a first order gradient descent and parameters were updated using Adam [19] during training. 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 network was trained on a single NVIDIA GeForce Titan X GPU. We use this network architecture to classify multiple categories of vocabulary datasets.

AURAL OUTPUT

Silent speech recognition of the AlterEgo system attempts to open up a unique opportunity to enable personalized bi-directional human-machine interfacing in a concealed and seamless manner, where the element of interaction is in natural language. 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 vocalized phrase 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. We use bone conduction headphones as the aural output, so as to not impede the user's sense of hearing.

WEARABLE FORM DESIGN

There are a number of design aspects that were considered with the intention of making the wearable system robust, and usable in a routine setting. Firstly, it is imperative for the electrodes to not shift position so as to maintain signal consistency. Secondly, it is desirable for the electrodes to maintain position on the target areas between multiple sessions of the user wearing and not wearing the device. Thirdly, while the electrodes must not move to stray forces, it is desirable for the positions to be adjusted to the positions of different users via one device that could be worn by multiple users. To that end, the form factor of the device is designed as a wearable that is worn around the back of the head, with extensions landing on the face to record signals from the afore-stated target areas (Figure 4). The band is 3D printed using photopolymer resin with a brass rod supporting it structurally, so as to maximize skin-electrode friction and minimize relative movement. The extensions are brass supports that provide rigidity to support electrodes while also being amenable to deliberate adjustment. Furthermore, the extensions are designed to attach to the frame in a modular manner such that specific extensions and electrodes could be pulled out for further experimentation.

APPLICATIONS

The AlterEgo system attempts to facilitate personal, discreet and seamless human-machine communication. In this section, we briefly present initial application explorations with the AlterEgo interface and demonstrate its utility as a personal cognition augmentation platform.

The current prototype implements modular neural networks in a hierarchical manner to accommodate for simultaneous accessibility to applications. The applications are initiated by internally vocalizing corresponding trigger words, for instance the word *IoT* for initiating wireless device control using the interface. At present, the vocabulary sets are modeled as *n*-gram sequences, where the recognition of a specific word assigns a probability distribution to subsequent vocabulary sets (Table 2).

The probability p_i can be assigned to a vocabulary set v_i based on previous recognition occurrences x_1 to x_{n-1} as $P(v_i|x_{n-1}...x_1) = p_i$. In the current setup, the probability $p_i = 1$ is assigned to vocabulary sets meant for specific applications, in a Markovian dependency arrangement, where each set is detected by a convolutional neural network. This hierarchy reduces the number of word possibilities to be detected within an application, thereby increasing the robustness of the current system.

The applications of AlterEgo can be classified under three broad categories:

Closed-loop interface

This scenario describes silent speech interface with a computing device where specific computer applications respond to internally vocalized queries through aural feedback, thereby enabling a closed-loop, silent and seamless conversation with a computing device. A few example implementations are described in this section.

The AlterEgo system allows the user to externalize any arithmetic expression to a computing device, through internally vocalizing the arithmetic expression and the computer subsequently relaying the computed value through aural feedback. For instance, the user could subvocalize the expression *2581 times 698 divide 2 add 13*, and the application would output the answer *900782* to the user, through bone conduction headphones. The device can be currently used to issue reminders and schedule tasks at specific times, which is aurally output to the user at corresponding times, thereby providing a form of memory augmentation to the user. The device also enables the user to access time using the interface, by silently communicating *world clock* and the name of a city, within a trained vocabulary set (Table 2).

Through such an interface, we explore if artificial intelligence (AI) could be democratized, and could instead act as an adjunct to human cognition in a personalized manner. As a demonstration of this, we implemented human-AI collaborative chess and Go through bi-directional silent speech, where the user would silently convey the game state and the AI would compute and then aurally output the next move to be played.

Open-loop interface

The interface could be used purely as an input modality to control devices and to avail services.

An example application is an IoT controller that enables the user to control home appliances and devices (switch on/off home lighting, television control, HVAC systems etc.) through internal speech, without any observable action. The interface can be personally trained to recognize phrases meant to access specific services. As an example, the internally vocalized phrase *Uber to home* could be used to book transport from the user's current destination using the interface. The interface could also be used as a silent input to Virtual Reality/Augmented Reality applications.

Human-human interface

The device also augments how people share and converse. In a meeting, the device could be used as a back-channel to silently communicate with another person. In the current instantiation of the device, the user can internally communicate 5 common conversational phrases to another person through the interface (Table 2). This could be naturally expanded with further user training.

We have created an environment for the device where applications could be developed catered to specific tasks. The environment asks the user to silently communicate the keywords of interest which is used for training for the application. In addition, the system potentially allows for peripheral devices to be directly interfaced with the system. For instance, lapel cameras and smart-glasses could directly communicate with the device and provide contextual information to and from the device.

EXPERIMENTAL EVALUATION

We sought to evaluate the word accuracy (WAcc) of the silent speech recognition model across multiple users, which formed our core experiment for the multi-user study. To evaluate the robustness of the platform, we recruited 10 participants (6 female) between 19 and 31 years old (μ = 23.6 years) to participate in experiments. None of the participants had any prior experience with the system evaluated in the study. We use the arithmetic computation application as our basis for accuracy evaluation.

Application Initializer	Vocabulary Set	
Arithmetic	0-9, multiply, add, subtract, divide, percent	0-9
IoT	light on, light off, fan on, fan off	
World Clock	Amsterdam, Beijing, Boston, Delhi, London, New York City, Johannesburg , Toronto	
Calendar	previous, next	
Chess	a-h, K, Q, R, N, 1-8	
Reply	0-9, finish	hello how are you, call you later, what’s up, yes, no

Table 2. Hierarchical organization of the vocabulary sets.

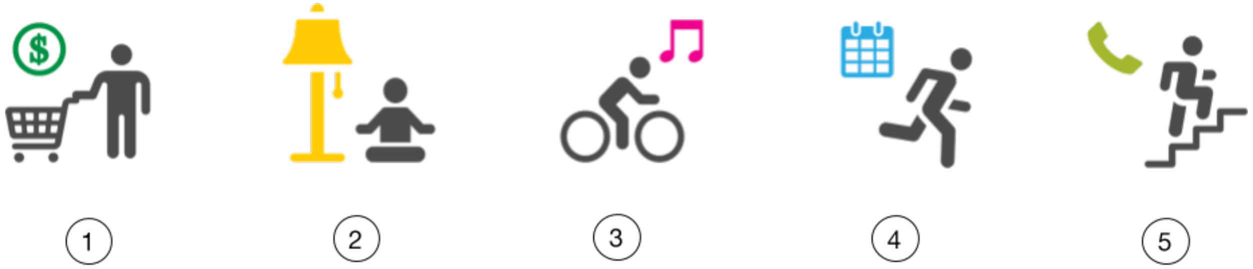

Figure 7. Examples of some applications or use cases: 1. Calculating totals during shopping (mathematics) 2. Controlling IoT devices 3. Controlling media 4. Temporal augmentation e.g. setting calendar meetings, current time, etc. 5. Responding to phone calls (Receive/reject)

The experiment was conducted in two phases.

First, we collected user silent speech data and evaluated word accuracy on a train test split. Second, we assessed the recognition latency of the interface by testing live inputs on the model, trained using the previous step.

In order to help the users understand silent speech, we showed the user a piece of text and asked the user to read it like (s)he silently read online articles, i.e. read to oneself and not out loud. For each participant, we showed them a total of 750 digits, randomly sequenced on a computer screen, and instructed the users to ‘read the number to themselves, without producing a sound and moving their lips’. The digits were randomly chosen from a total of 10 digits (0 to 9), such that each digit exactly appeared 75 times. The data was recorded for each user with the trigger voltages marking word-label alignments.

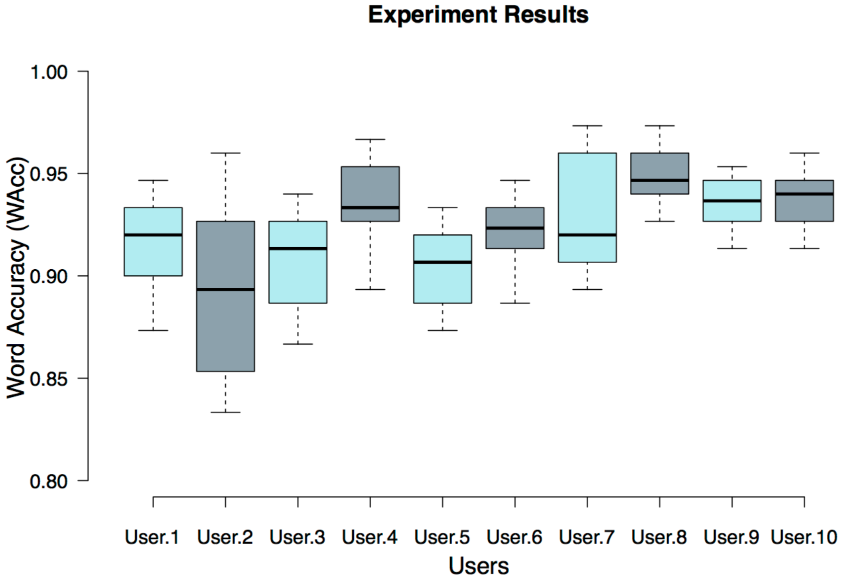

Figure 8. Word accuracy (WAcc) of the silent speech recognition model evaluated on 10 users, using 10 training testing splits of the arithmetic computation dataset. The edges of boxes represent 1st and 3rd quartile respectively, and the whiskers extend to 100% coverage.

Quantitative results

The user data was split according to an 80/20 random split for training and testing for each user. The word accuracies for each user were recorded for 10 runs of training and testing. Figure 8 shows the word accuracies for each user and each run. The average accuracy over all runs for all the users 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 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).

DISCUSSION

The results from our preliminary experiments show that the accuracy of our silent speech system is at par with the reported word accuracies of state-of-the-art speech recognition systems, in terms of being robust enough to be deployed as voice interfaces, albeit on smaller vocabulary sets. The promising results from the current interface show that AlterEgo could be a step forward in the direction of human-machine symbiosis. We plan to conduct further experiments to test the system for an augmented vocabulary dataset in real-world ambulatory settings.

The concept of human-machine symbiosis has been suggested by several, such as Engelbart [7,8], Licklider [20] and Ashby [2,3] to propose the combination of human and machine intelligence as a goal to be pursued and as an effective computational strategy as opposed to either human or machine intelligence acting independently. Recently, there have been several advances in the area of AI/machine intelligence, prompting anxiety with respect to consequences for human labor and societies [11]. The current viewpoint commonly places machine and human intelligence at odds. Through the AlterEgo device, we seek to move in the step to couple human and machine intelligence in a complimentary symbiosis. As smart machines work in close unison with humans, through such platforms, we anticipate the progress in machine intelligence research to complement intelligence augmentation (IA) efforts, which would lead to an eventual convergence - to augment humans in wide variety of everyday tasks, ranging from computations to creativity to leisure.

FUTURE WORK

There remain many avenues for future work. In particular, we identify the following key future tasks for our silent speech device:

1. *Collect more data to develop a more generalized multi-user silent speech recognition model:* We aim to develop a generalized multi-user system that is user-independent, but can also be tuned and personalized for each user when they start using the device.

2. *Extend the system to include a broader vocabulary of words:* In the current instantiation, we implemented accessibility to multiple vocabulary sets simultaneously, albeit on limited data. Our experimental evaluation was based on an arithmetic computation application. We plan to augment our recognition models to accommodate for a

larger dataset, and plan to follow this with thorough multi-user longitudinal accuracy tests of our system.

3. *Test the system in real-world ambulatory settings:* Our existing study was conducted in a stationary setup. In the future, we would like to conduct longitudinal usability tests in daily scenarios.

CONCLUSION

Silent speech entails that the user communicates with the device by internally talking to oneself instead of actual speech articulation. We akin this to reading something to oneself without moving one's lips, producing an audible sound and without any discernable action.

Silent speech interfaces allow the user to communicate with computers, applications and people as seamlessly as they do through speech interfaces (telecommunications devices, speech based smart assistants, social robots etc.), but without the overhead of saying things out loud. As a result, silent speech interfaces are more private and personal for each user, and do not conflict with the existing verbal communication channels between people. We envision that the usage of our device will interweave human and machine intelligence to enable a more natural human-machine symbiosis that extends and augments human intelligence and capability in everyday lives.

ACKNOWLEDGEMENTS

The authors would like to thank Neo Mohsenvand, Mina Khan, Ashris Choudhury, Jason Robinson for their valuable inputs and feedback on the manuscript.

REFERENCES

1. Dario Amodei, Rishita Anubhai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Jingdong Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Erich Elsen, Jesse Engel, Linxi Fan, Christopher Fougner, Tony Han, Awni Hannun, Billy Jun, Patrick LeGresley, Libby Lin, Sharan Narang, Andrew Ng, Sherjil Ozair, Ryan Prenger, Jonathan Raiman, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Yi Wang, Zhiqian Wang, Chong Wang, Bo Xiao, Dani Yogatama, Jun Zhan, and Zhenyao Zhu. 2015. Deep Speech 2: End-to-End Speech Recognition in English and Mandarin. *arXiv [cs.CL]*. Retrieved from http://arxiv.org/abs/1512.02595
2. W. Ross Ashby. 1956. Design for an intelligence-amplifier. *Automata studies* 400: 215–233.
3. W. Ross Ashby. 1957. An introduction to cybernetics. Retrieved from http://dspace.utalca.cl/handle/1950/6344
4. Alan Baddeley, Marge Eldridge, and Vivien Lewis. 1981. The role of subvocalisation in reading. *The Quarterly Journal of Experimental Psychology Section A* 33, 4: 439–454.
5. Richard A. Bolt. 1980. &Ldquo;Put-that-there&Rdquo;: Voice and Gesture at the Graphics Interface. *SIGGRAPH Comput. Graph.* 14, 3: 262–270.
6. Jonathan S. Brumberg, Alfonso Nieto-Castanon, Philip R. Kennedy, and Frank H. Guenther. 2010. Brain-Computer Interfaces for Speech Communication. *Speech communication* 52, 4: 367–379.
7. Douglas C. Engelbart. 2001. Augmenting human intellect: a conceptual framework (1962). *PACKER, Randall and JORDAN, Ken. Multimedia. From Wagner to Virtual Reality. New York: WW Norton & Company*: 64–90.
8. Douglas C. Engelbart and William K. English. 1968. A Research Center for Augmenting Human Intellect. In *Proceedings of the December 9-11, 1968, Fall Joint Computer Conference, Part I* (AFIPS '68 (Fall, part I)), 395–410.
9. M. J. Fagan, S. R. Ell, J. M. Gilbert, E. Sarrazin, and P. M. Chapman. 2008. Development of a (silent) speech recognition system for patients following laryngectomy. *Medical engineering & physics* 30, 4: 419–425.
10. Victoria M. Florescu, Lise Crevier-Buchman, Bruce Denby, Thomas Hueber, Antonia Colazo-Simon, Claire Pillot-Loiseau, Pierre Roussel, Cédric Gendrot, and Sophie Quattrocchi. 2010. Silent vs vocalized articulation for a portable ultrasound-based silent speech interface. In *Eleventh Annual Conference of the International Speech Communication Association*. Retrieved from http://www.gipsa-lab.inpg.fr/~thomas.hueber/mes_documents/florescu_etal_interspeech_2010.PDF
11. Carl Benedikt Frey and Michael A. Osborne. 2017. The future of employment: How susceptible are jobs to computerisation? *Technological forecasting and social change* 114, Supplement C: 254–280.
12. A. Graves, A. r. Mohamed, and G. Hinton. 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 6645–6649.
13. Jefferson Y. Han. 2005. Low-cost Multi-touch Sensing Through Frustrated Total Internal Reflection. In *Proceedings of the 18th Annual ACM Symposium on User Interface Software and*

Technology (UIST ’05), 115–118.

14. William J. Hardcastle. 1976. *Physiology of speech production: an introduction for speech scientists*. Academic Press.
15. Tatsuya Hirahara, Makoto Otani, Shota Shimizu, Tomoki Toda, Keigo Nakamura, Yoshitaka Nakajima, and Kiyohiro Shikano. 2010. Silent-speech enhancement using body-conducted vocal-tract resonance signals. *Speech communication* 52, 4: 301–313.
16. Robin Hofe, Stephen R. Ell, Michael J. Fagan, James M. Gilbert, Phil D. Green, Roger K. Moore, and Sergey I. Rybchenko. 2013. Small-vocabulary speech recognition using a silent speech interface based on magnetic sensing. *Speech communication* 55, 1: 22–32.
17. Thomas Hueber, Gérard Chollet, Bruce Denby, and Maureen Stone. 2008. Acquisition of ultrasound, video and acoustic speech data for a silent-speech interface application. *Proc. of ISSP*: 365–369.
18. Jorgensen, C., & Binsted, K. (2005, January). Web browser control using EMG based sub vocal speech recognition. In *System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on* (pp. 294c-294c). IEEE.
19. Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *arXiv [cs.LG]*. Retrieved from http://arxiv.org/abs/1412.6980
20. J. C. R. Licklider. 1960. Man-Computer Symbiosis. *IRE Transactions on Human Factors in Electronics* HFE-1, 1: 4–11.
21. S. Mitra and T. Acharya. 2007. Gesture Recognition: A Survey. *IEEE transactions on systems, man and cybernetics. Part C, Applications and reviews: a publication of the IEEE Systems, Man, and Cybernetics Society* 37, 3: 311–324.
22. A. Nijholt, D. Tan, G. Pfurtscheller, C. Brunner, J. d. R. Millán, B. Allison, B. Graimann, F. Popescu, B. Blankertz, and K. R. Müller. 2008. Brain-Computer Interfacing for Intelligent Systems. *IEEE intelligent systems* 23, 3: 72–79.
23. Sharon Oviatt, Phil Cohen, Lizhong Wu, Lisbeth Duncan, Bernhard Suhm, Josh Bers, Thomas Holzman, Terry Winograd, James Landay, Jim Larson, and David Ferro. 2000. Designing the User Interface for Multimodal Speech and Pen-Based Gesture Applications: State-of-the-Art Systems and Future Research Directions. *Human–Computer Interaction* 15, 4: 263–322.
24. Anne Porbadnigk, Marek Wester, Jan-P Calliess, and Tanja Schultz. 2009. EEG-BASED SPEECH RECOGNITION Impact of Temporal Effects.
25. Michael Wand and Tanja Schultz. 2011. Session-independent EMG-based Speech Recognition. In *Biosignals*, 295–300.
26. M. Wand, J. Koutník, and J. Schmidhuber. 2016. Lipreading with long short-term memory. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6115–6119.
27. Nicole Yankelovich, Gina-Anne Levow, and Matt Marx. 1995. Designing SpeechActs: Issues in Speech User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI ’95), 369–376.
28. iOS - Siri. *Apple*. Retrieved October 9, 2017 from https://www.apple.com/ios/siri/
29. Alexa. Retrieved from https://developer.amazon.com/alexa
30. Cortana | Your Intelligent Virtual & Personal Assistant | Microsoft. Retrieved October 9, 2017 from https://www.microsoft.com/en-us/windows/cortana
31. Google Home. *Google Store*. Retrieved October 9, 2017 from https://store.google.com/us/product/google_home?hl=en-US
32. Echo. Retrieved from https://www.amazon.com/Amazon-Echo-And-Alexa-Devices/b?ie=UTF8&node=9818047011