Analysis into the use of EMG & AI to unlock a voice for the speech-impaired

Jacob van Steyn, Luke Savoie, Ali Yousefi

Computer Science Dept. & Data Science Dept., Worcester Polytechnic Institute
Worcester, USA

jbvansteyn@wpi.edu

lrsavoie@wpi.edu

***Abstract*— Here, we demonstrate the use of AI to unlock a 'voice' for the speech impaired, creating a novel way for people to speak silently and imperceptibly when in public or otherwise, opening a new medium of communication for people with speech disorders. By simply thinking of a number in your mind, our system can detect with high accuracy what your were planning to say within a low latency. Minute muscular activation signals occur when a person intends to speak, which we capture in real-time and map to the numbers. We argue our machine learning pipeline is robust and repeatable, and it reaches a classification accuracy better than 99% for the numbers 1-5 and provide a reproducible framework for further research.**

***Keywords*—Subvocalization, Biosensing, EMG, Human-Computer Interaction, Machine Learning**

I. Introduction

The concept of "mind reading" has long been popularized in mainstream media as a futuristic technology. Although research in this field is scarce, subvocalization has promise to allow for a means to decode the thoughts or intentions of an individual.

Subvocalization is the process of silently speaking to oneself without physically moving any muscles, vocal cords, or speaking aloud – simply thinking of a word intentionally. During this process, the muscles of the vocal cords move imperceptibly, causing small electrical signals to be generated as the body prepares to speak the word which is being thought [1,2]. These electrical signals can be recorded with the use of electromyography (EMG) sensors placed in proximity to the applicable muscles.

The ability to decode subvocalized words has a great potential to assist individuals with speech impairments (e.g. people with vocal cord paralysis, aphonia, ALS, cancer, stroke, etc.). Additionally, it can enable seamless human-computer interaction and silent communication between people. However, current technologies for subvocalization are generally limited in their capabilities or complex and invasive, highlighting the need for further advancements in this area.

This paper expands upon the limited research and covers three primary topics:

- Collecting EMG biosignals and important considerations
- Optimization of sensor placement for subvocalization
- Signal processing and using machine learning to decode unspoken words

We aim to demonstrate the viability of using AI to decipher silent speech and create a framework for future research and development in this area by presenting our results and findings on the topic.

II. Literature Review

The concept of subvocalization has been of interest to researchers for decades. Recent advances in technology have enabled researchers to explore the potential of subvocalization in areas such as human-computer interaction and speech therapy.

There are several related areas of research that precede the concept of subvocalization, yet maintain some of the core concepts and application. Some examples we came across in our research were eye tracking communication devices, read aloud technology, movement controlled devices, and general Augmentative and Alternative Communication (AAC) devices. AAC refers to technology that enables communication for individuals with physical or complex communication impairments. Over the past three

decades, AAC has expanded from interpersonal communication to access of information and services over the internet [3]. However, the limited use and abandonment of AAC technologies remains high.

Eye-tracking technology allows individuals with limited mobility and complex communication needs to control devices such as augmentative and alternative communication devices, virtual reality systems, and video games using their eye movements [4]. In the ICU, eye-tracking devices have been found to enhance psychosocial status, communication ability, and reduce delirium among patients [5]. Further exploration is needed to understand the limitations and benefits of the technology, but eye-tracking technology has been shown to make a significant impact on the lives of individuals with complex communication needs and has the potential to improve outcomes in various populations [6].

Read aloud devices, which utilize a terminology called text-to-speech (TTS), is a form of technology that converts written text into spoken words [7]. ARTIC [8], a Czech TTS system, is one example of TTS technology we came across in research.. This type of AAC technology is being widely implemented to assist students' reading comprehension skills, with studies indicating a potentially positive effect on reading comprehension for students with reading difficulties.

Work has also been done exploring teeth-clenching as a target selection mechanism in AR applications. The creators of ClenchClick developed a teeth-clenching detection system [9] and evaluated its performance in target selection tasks compared to two baseline methods. ClenchClick outperformed the other methods in workload, physical load, accuracy, and speed.

Technologies we have discussed so far have mostly been used in a professional setting of helping a certain group of people. There is also an established market and use case for vocal recognition technology such as Google Home, Amazon Echo, Siri, and Cortana. These systems have been developed across multiple decades and provide a pathway for subvocalized speech recognition technologies to become usable and marketable.

Prior studies on human-computer interaction have focused primarily on physical or vocal input. However, research on non-physical human-computer interaction through subvocalization has been limited and varies in scope and methodology. Two primary approaches have been studied: invasive and non-invasive systems. Invasive systems are typically surgically implanted and permanent, making them complex and expensive solutions. Thus, our study focused on non-invasive alternatives. One such approach was conducted by a research team at MIT, who used a limited set of words and demonstrated accurate results in classification [10]. Other similar approaches simplified the problem to determining complexity and focus [11]. We aimed to build on their findings, but found several elements of their research to be unclear and difficult to reproduce.

III. EMG Biosensing

EMG (Electromyography) Biosensing is a technique used to detect and measure the electrical activity generated by nervous impulses which signal for the contraction of skeletal muscles. In the context of speech production, the electrical signals generated by the muscles involved in facial articulation can be measured and recorded using surface EMG sensors (sEMG) placed on the skin above these muscles. When performing subvocalization, muscles in the face and neck create myoelectrical signals in preparation for speech production. These electrical impulses generate a time-varying potential difference pattern over various muscles which can be captured by sEMG sensors. We chose to use electromyography sensors in our research using pre-gelled Ag/Cl adhesive electrodes. These surface adhered electrodes provide a non-invasive method to capture the subtle muscle activity associated with subvocalization.

To determine optimal sensor placement, we conducted multiple rounds of data collection and analysis in order to find muscle locations which would provide the lowest Signal-to-Noise (SNR) ratio for our dataset. Signal-to-Noise Ratio is a measure of the strength of a signal relative to the background noise present in the recording. In the context of sEMG biosensing, SNR is an important metric because it indicates the reliability of the myoelectric signal being recorded. The higher the SNR, the easier it is to distinguish the subtle electrical activity associated with speech production from background noise. Since the electrical signals we're trying to detect are very minute and susceptible to various types of noise, it's crucial to find the optimal sensor placement to obtain the highest possible SNR in our data. By selecting muscle locations with a low SNR, we can ensure that our recorded signals are as clear and reliable as possible, which will improve the accuracy of our machine learning model in decoding subvocalized words.

In order to find optimal sensor placement we determined a few possible locations using the expertise of Professor Adam Lammert, who specializes in speech pathology and biomechanics. Including facial and throat muscle structure, we narrowed our focus down to six primary locations:

Region	Common Location
Laryngeal	Throat
Mentalis	Chin
Hyoid	Under Chin
Buccal	Mouth
Orbital	Cheek
Masseter	Jaw

Fig. 1 - Sensor Placement Locations

Recordings from various configurations of these locations were collected and analyzed in order to select final locations.

To collect data, participants were instructed to silently subvocalize numbers from one to five repeatedly in their heads for a period of five minutes per number while wearing the sEMG sensors. An accelerometer was attached to the participant's index finger, which they would tap on the table when thinking of a word. The accelerometer data was used to segment the continuous stream of data into individual samples of data. The participants were given a rest period of five seconds in between each subvocalization task.

IV. Data Processing & Analysis

In order to analyze the quality of our data, we first split our collected data into samples, then further subdivided the samples by region, number, and participant. The primary goal of our analysis was to find sensor locations which provided quality data and easily discernible differences between our target classes.

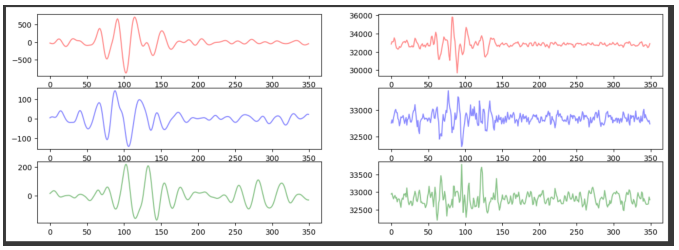

Figure 2 - Filtered vs Unfiltered sEMG Signals

After collecting our data, we performed a preprocessing step to reduce and eliminate as much interference as possible before analyzing our samples. In order to eliminate cardiac and other biological interference, we placed a reference electrode behind the right ear. We used a Butterworth filter to reduce electrical line interference from nearby power sources, which we found to be a common source of noise in our recorded signals. The filter has a passband of 2 to 45Hz, which means it allows frequencies in that range to pass through while attenuating frequencies outside of that range. While the filter effectively reduced noise in our signals, it did not completely eliminate all 60Hz interference. To further reduce the impact of 60Hz interference, we applied an

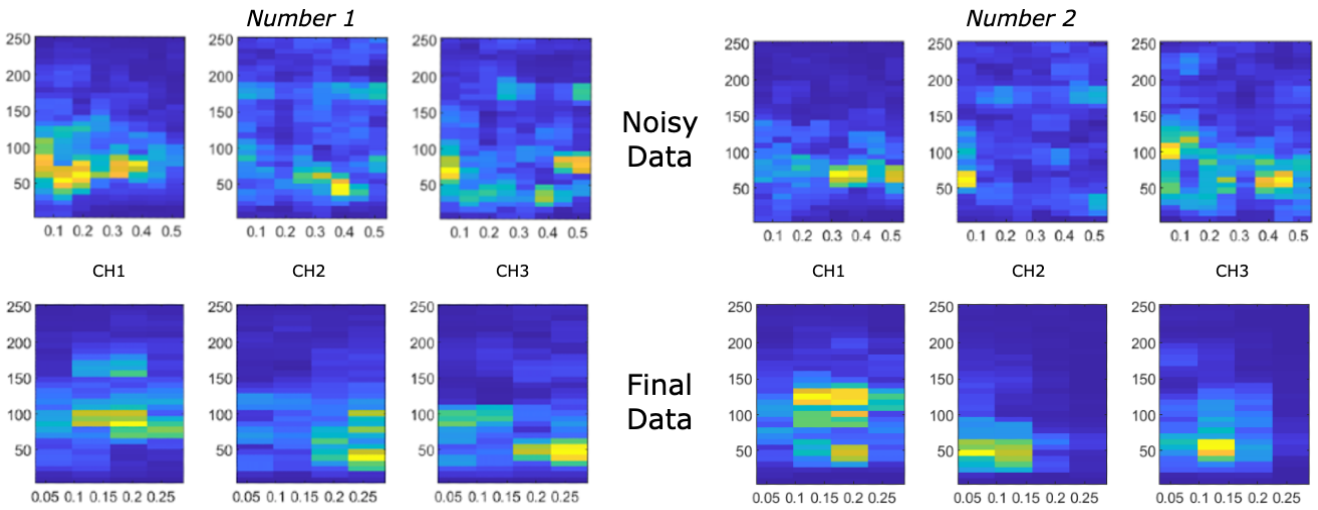

Figure 3 - Noisy vs Final data (3 CH)

additional Infinite Impulse Response (IIR) notch filter. The IIR notch filter is a type of filter that can be used to selectively remove a narrow range of frequencies, in this case, 60Hz, from the signal while leaving other frequencies largely unaffected. With these preprocessing steps in place, we were able to obtain high-quality signals from our sEMG sensors and reduce the noise in our data (Fig. 2).

To evaluate the accuracy of our collected data, we chose to visually analyze samples using a spectrogram representation. Samples were grouped by number and location, then for each channel all of the samples in that subset were averaged together.

The preprocessed data was then transformed into a spectrogram representation using a multi-tapered method with a window length of 128 ms and 50% overlap. The spectrograms were then plotted for each channel (out of three) per number and location. This spectral analysis allowed us to visually evaluate the quality of our data.

Due to the averaging of samples we can see noise when recording poor quality data and distinct clusters with useful data (Fig. 3). Using this technique, we chose the three best performing locations for which to develop our model with.

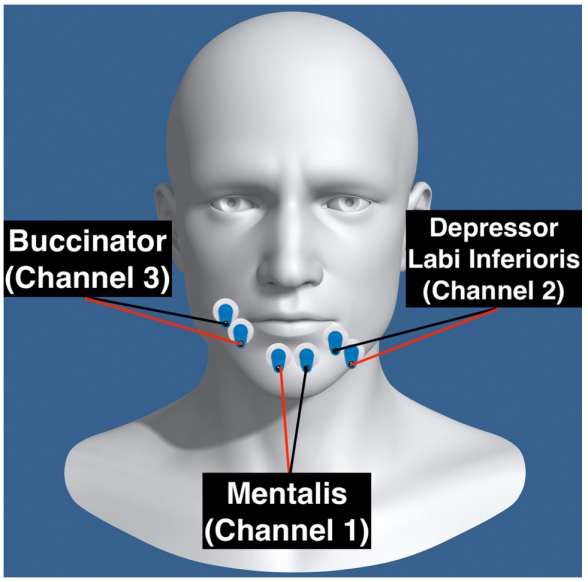

Figure 4 - sEMG Sensor Locations

Using spectral analysis we determined three final sensor regions located on the Mentalis, Depressor Labi Inferioris, and Buccinator (Fig. 4).

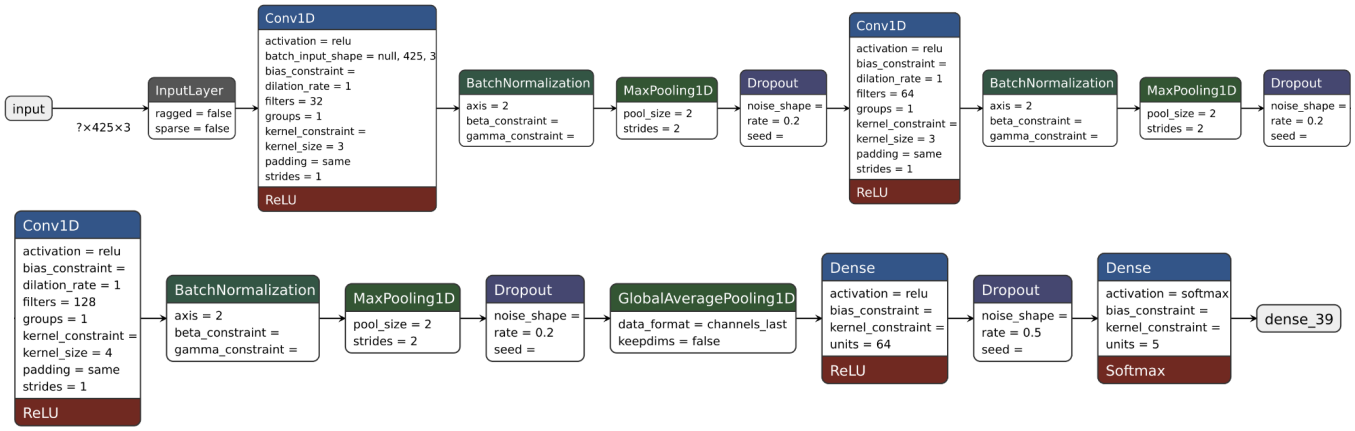

Figure 5 - Final CNN Model Architecture

V. Neural Network

In an attempt to reduce the time-shifting bias incurred from signalling subvocalization through accelerometer taps, we augmented preprocessed training data by shifting samples by up to 25 ms in either direction from the signal onset. By doing so, we increased the size of our dataset, thereby providing more varied examples for training, which allowed us to further improve our model's test accuracy. Ultimately, the model was trained on a dataset of 10,800 samples, divided into five classes, each consisting of 425ms of sEMG data with 3 channels per sample. 6,100 unshifted samples were used to validate the accuracy of the model with 3,600 samples used during training and 2,600 samples kept entirely separate to test the model on after training completed.

To classify our subvocalized data on the numbers 1-5, we tried a cohort of differing machine learning processes. After experimenting with various neural network types, we discovered that a Convolutional Neural Network (CNN) was the most effective at classifying our time-series data. This was primarily due to the fact that CNN's are particularly adept at detecting features in data with a temporal or spatial nature, which is crucial for speech recognition. We found that a CNN model performed exceptionally well with our dataset, and was able to accurately classify our data.

However, we did not solely rely on the CNN model for our analysis. In an effort to identify the most suitable structure for our classification task, we explored other neural network architectures. We tested several RNN models, which are designed to process sequential data, as well as Long Short-Term Memory (LSTM) models, a type of recurrent neural network that can learn long-term dependencies. Despite the promise of these models, they appeared to underperform in comparison to the CNN.

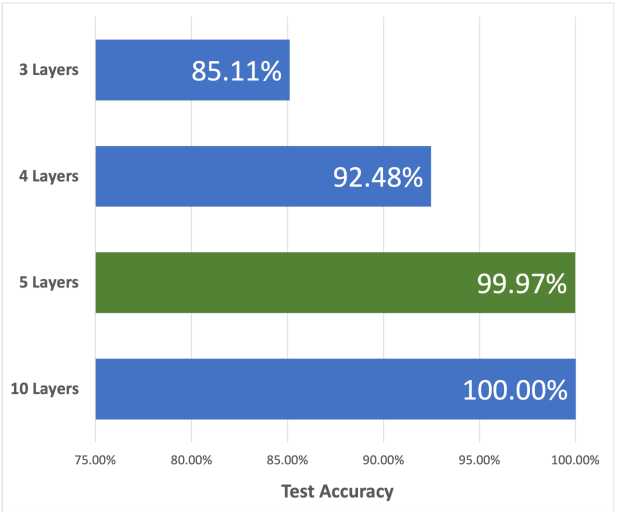

Figure 6 - Model Complexity (CNN+Dense Layers) vs. Test Accuracy

In addition to experimenting with different neural network types, we also varied the complexity of our

CNN model to evaluate its impact on classification performance. This included changing the structures of the model, such as adjusting filter sizes, the number of layers, and using different pooling techniques. We also adjusted the number of neurons in the hidden layers, the learning rate, and the regularization parameters to minimize the model complexity while still retaining a high classification accuracy. The final model structure that we settled on minimized model complexity, allowing for quicker inference times, which is an important aspect in speech interfaces. Figure 6 illustrates the impact of varying model complexity on classification performance, and Figure 5 shows the final model architecture.

VI. Results

Our results showed that the final trained model was able to accurately classify the numbers 1-5 from facial subvocalization signals with an average accuracy of 99.9%. The confusion matrix for the classification results is shown in Figure 7.

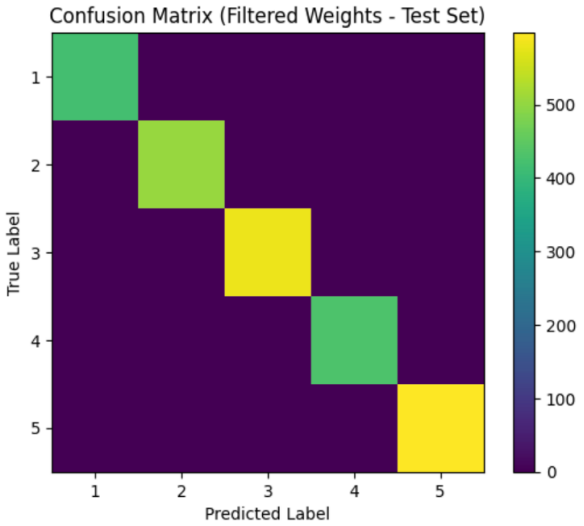

Figure 7 - Confusion matrix for the classification of numbers 1-5

Comparison with previous research showed significant improvements in accuracy rates. Previous studies using machine learning algorithms to classify subvocal speech reported accuracy rates around 95%. Again, this study had inconsistencies and a non reproducible framework. This study also required 8 sEMG sensors, while our data was recorded with 3.

The choice of which ML technique was used also strongly affected our results. When experimenting across various ML models, we found that though model architecture affects accuracy – data quality was significantly more important. Depending on the network architecture and data quality, our accuracy varied from as 30% to 60% when using poor quality data. After prioritizing the optimization of our sensor placement and model architecture, we started to see significant improvements in classification accuracy.

In our evaluation of the trained model, we performed testing on 2,600 samples of previously unseen data. The results demonstrated a remarkable level of accuracy, with the model able to perfectly classify a subset of the data with 100% accuracy. In order to ensure that the accuracy was not the result of overfitting or test data leakage into the training data, we undertook several steps.

Firstly, we experimented with varying model complexities to ensure that the model was not overfitting on the training data. We then conducted a thorough analysis of the training process to ensure that there was no data leakage between the test and training sets. Through these measures, we were able to establish the accuracy of the model with confidence.

Further testing of different model structures with varying complexities revealed a strong correlation between the complexity of the model and its performance (Fig. 6).

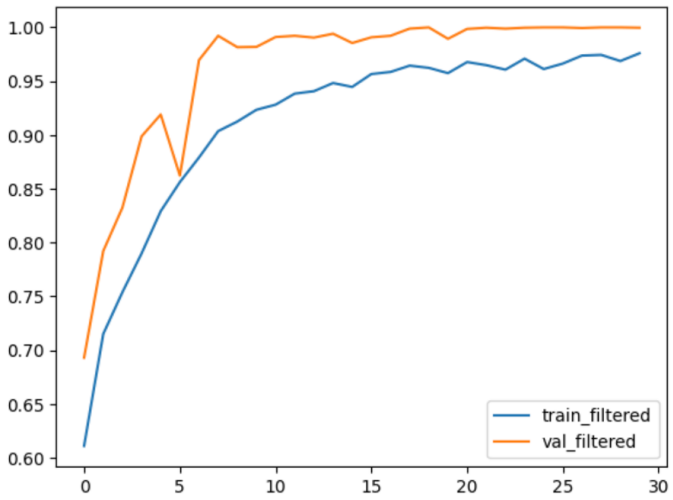

Figure 8 - Final model training & validation Accuracy per epoch

We demonstrate the successful application of machine learning techniques for accurate classification of subvocalized speech signals from facial muscles. Our final trained model achieved an accuracy of 99.9% in classifying numbers 1-5, outperforming previous studies that relied on more sensors. Our findings also emphasize the importance of optimizing both data quality and model architecture for achieving high classification accuracy. The level of accuracy achieved by this model demonstrates its robustness and further potential for additional research.

VII. Discussion

We will provide a framework for further development in the field of subvocalization and human-computer interaction, particularly for individuals with speech impairments. A future version of this system could make communication for such individuals more efficient and convenient, thereby enhancing their quality of life. Additionally, this technology could also be useful in scenarios where silent communication is necessary, such as in loud environments or when maintaining secrecy is critical.

As we chose to categorize only upon numbers, more research and data would be required to recognize a broader range of words or phrases. Moreover, the sample size of our participants was fairly limited, underscoring the need for future studies with a larger sample size to validate the ability to generalize our findings across a wide range of different individuals. When investigating generalization, attempting to use transfer learning may enable more accurate and individualized models without the need for large data collection on a person-by-person basis [12]. In addition, more advanced or specific AI techniques may be necessary to develop in order to accurately interpret sEMG signals during subvocalization, particularly in environments with high electrical interference.

The opportunities for the practical application of this system are numerous, and could be utilized in a number of different industries. For instance, the technology could be incorporated into devices such as glasses, helmets, “smart stickers” [13], or other applications. This integration would allow for hands-free and simple communication between users and technology. Silent and privacy focused communication has applications in noisy environments, situations which require secrecy, or even simply to allow people to interact with their technology without being noticed. Industries such as gaming could also utilize the system to provide users with a more realistic and lifelike gameplay.

Despite the potential applications, there are numerous ethical concerns related to the development and use of this technology that need to be addressed [14]. It is essential to consider the potential privacy implications of recording and interpreting an individual's “thoughts”. Additionally, it is important that this technology is accessible to all individuals by ensuring further data collection has a broad representation from various groups of people.

VIII. Conclusion

Our proposed system demonstrated promising results in accurately interpreting sEMG signals during subvocalization. Although there are limitations to the study, the findings have significant implications for the field of subvocalization and human-computer interaction.

Our results show that using our methodology classification of subvocalized thoughts can be performed with an extremely high accuracy, demonstrating the potential for the application of decoding subvocalized speech, and providing a basis for further exploration.

Acknowledgment

We would like to thank Professor Ali Yousefi for providing invaluable insights, assisting with visualization & signal analysis, and advising our research. Additionally, we'd like to thank Professor Adam Lammert for providing critical information about speech production & what muscles are involved to focus on.

References

[1] Helou, L. B., Welch, B., Wang, W., Rosen, C. A., & Verdolini Abbott, K. (2021). Intrinsic Laryngeal Muscle Activity During Subvocalization. Journal of Voice. https://doi.org/10.1016/j.jvoice.2021.01.021

[2] Aarons, L. (1971). Subvocalization: Aural and Emg Feedback in Reading. Perceptual and Motor Skills, 33(1), 271–306. https://doi.org/10.2466/pms.1971.33.1.271

[3] Shane, H. C., Blackstone, S., Vanderheiden, G., Williams, M., & DeRuyter, F. (2012). Using AAC Technology to Access the World. Assistive Technology, 24(1), 3–13. https://doi.org/10.1080/10400435.2011.648716

[4] Kaufman, A. E., Bandopadhay, A., & Shaviv, B. D. (1993). An eye tracking computer user interface. Proceedings of 1993 IEEE Research Properties in Virtual Reality Symposium, 120–121. https://doi.org/10.1109/VRAIS.1993.378254

[5] Garry, J., Casey, K., Cole, T. K., Regensburg, A., McElroy, C., Schneider, E., Efron, D., & Chi, A. (2016). A pilot study of eye-tracking devices in intensive care. Surgery, 159(3), 938–944. https://doi.org/10.1016/j.surg.2015.08.012

[6] Lariviere, J. A. (2015). Eye Tracking: Eye-Gaze Technology. In I. Söderback (Ed.), International Handbook of Occupational Therapy Interventions (pp. 339–362). Springer International Publishing. https://doi.org/10.1007/978-3-319-08141-0_23

[7] Acero, A. (2000). An overview of text-to-speech synthesis. 2000 IEEE Workshop on Speech Coding. Proceedings. Meeting the Challenges of the New Millennium (Cat. No.00EX421), 1-. https://doi.org/10.1109/SCFT.2000.878372

[8] Tihelka, D., Hanzlíček, Z., Jůzová, M., Vít, J., Matoušek, J., & Grůber, M. (2018). Current State of Text-to-Speech System ARTIC: A Decade of Research on the Field of Speech Technologies. In P. Sojka, A. Horák, I. Kopeček, & K. Pala (Eds.), Text, Speech, and Dialogue (pp. 369–378). Springer International Publishing. https://doi.org/10.1007/978-3-030-00794-2_40

[9] Shen, X., Yan, Y., Yu, C., & Shi, Y. (2022). ClenchClick: Hands-Free Target Selection Method Leveraging Teeth-Clench for Augmented Reality. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 6(3), 139:1-139:26. https://doi.org/10.1145/3550327

[10] Kapur, A., Sarawgi, U., Wadkins, E., Wu, M., Hollenstein, N., & Maes, P. (2020). Non-Invasive Silent Speech Recognition in Multiple Sclerosis with Dysphonia. Proceedings of the Machine Learning for Health NeurIPS Workshop, 25–38. https://proceedings.mlr.press/v116/kapur20a.html

[11] Parnin, C. (2011). Subvocalization—Toward Hearing the Inner Thoughts of Developers. 197–200. https://doi.org/10.1109/ICPC.2011.49

[12] Gupta, J., Pathak, S., & Kumar, G. (2022). Deep Learning (CNN) and Transfer Learning: A Review. Journal of Physics: Conference Series, 2273(1), 012029. https://doi.org/10.1088/1742-6596/2273/1/012029

[13] Lopes, P. A., Vaz Gomes, D., Green Marques, D., Faia, P., Góis, J., Patrício, T. F., Coelho, J., Serra, A., de Almeida, A. T., Majidi, C., & Tavakoli, M. (2019). Soft Bioelectronic Stickers: Selection and Evaluation of Skin-Interfacing Electrodes. Advanced Healthcare Materials, 8(15), 1900234. https://doi.org/10.1002/adhm.201900234

[14] Jwa, A. S., & Poldrack, R. A. (n.d.). Addressing privacy risk in neuroscience data: From data protection to harm prevention.