Voice Restoration Device Using Machine Learning of Acoustic and Visual Output During Electrolarynx Use

(Technical Paper)

A Review of Algorithmic Bias in the American Healthcare System

(STS Paper)

A Thesis Prospectus Submitted to the

Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements of the Degree

Bachelor of Science, School of Engineering

Katherine M. Taylor

Fall, 2021

Technical Project Team Members

Sameer Agrawal

Surabhi Ghatti

Medhini Rachamallu

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Katherine M. Taylor

James J. Daniero, M.D., Department of Otolaryngology – Head and Neck Surgery

Haibo Dong, Ph.D., Department of Mechanical and Aerospace Engineering

Rider Foley, Department of Engineering and Society

Introduction

Every year, over 3000 patients in the United States alone receive a total laryngectomy, which entails the removal of the vocal cords, due to laryngeal cancer (Kohlberg, Gal, & Lalwani, 2016). Recipients of this surgery, called laryngectomees, lose the ability to speak and require alternative means of communication. Any therapy involving voice restoration is preferred over a non-verbal communication strategy since the ability to speak is associated with a higher quality of life (Antin, Breheret, Goineau, Capitain, & Laccourreye, 2021; Wulff, Højager, Wessel, Dalton, & Homøe, 2021). One common voice restoration therapy is the electrolarynx, a device placed under the chin. The electrolarynx emits vibrations that are transmitted through the skin to the throat; the vibrations are shaped into words using the lips, tongue, and teeth. The electrolarynx is preferred over other forms of voice restoration therapies because it is non-invasive, has no complications, and is most cost effective (Carr, Schmidbauer, Majaess, & Smith, 2000).

However, the electrolarynx poses several concrete problems that need to be addressed. Despite high self-evaluation scores for basic functioning, users of all voice restoration therapies report a quality of life that is lower than average, particularly in the areas of anxiety and depression. Laryngectomees are also plagued by an inability to communicate in noisy settings and thus demonstrate a higher level of social avoidance than normal (Carr et al., 2000; Cox & Doyle, 2014). This is furthered by the presence of generic vibrational noise that can drown out the voice of the speaker (T. Masaki, personal communication, October 20, 2021; Padmini, 2017). The electrolarynx requires fewer instructional speech therapist appointments than other voice restoration therapies but still takes time and effort to learn to use effectively (Carr et al., 2000). Finally, it is important to remember that laryngectomees have just made the decision between losing their voice and losing their life, which is an extremely difficult setback to overcome mentally (Gates et al., 1982; T. Masaki, personal communication, October 20, 2021). Thus, an ideal voice restoration therapy that combines the important values of non-invasiveness and low cost with higher speech intelligibility and ease of use is essential to improve the quality of life of laryngectomees.

The proposed solution to this ideal voice restoration therapy entails the creation of an artificial neural network (ANN) that translates both lip movements and electrolarynx audio output into computer-generated speech. Though this intends to reduce the number of problems associated with electrolarynx use, the use of machine learning extends the possibility of bias in the speech generation algorithm. Multiple instances of a supposedly neutral machine learning algorithm introducing bias into a judgment have been observed in society (Martin, 2019; Mehrabi, Morstatter, Saxena, Lerman, & Galstyan, 2021; Yates, Gulati, & Weiss, 2013). As this algorithm could potentially impact the quality of life and mental health of the user, it is of utmost importance that potential biases are identified and mitigated before the final product is released. After a thorough description of the significance, innovation, and intended methods of the ANN, the need for bias identification and mitigation in machine learning algorithms will be explored.

Voice Restoration Device Using Machine Learning of Acoustic and Visual Output During Electrolarynx Use

The use of ANNs rather than traditional machine learning algorithms to solve problems has become popular due to increased accuracy and user simplicity (O'Mahony et al., 2020). ANNs are modeled after the connections between neurons in the human brain and are comprised of input and output layers as well as hidden layers that produce the final output. ANNs improve accuracy through iteration through a training dataset while avoiding overfitting by evaluating the resulting model on a testing dataset (Kohlberg et al., 2016). Current research in artificial lip reading focuses on the implementation of ANNs to analyze different points on the lip, which can achieve higher accuracy than human lip reading (Assael, Shillingford, Whiteson, & de Freitas, 2016; Kohlberg et al., 2016). Current efforts to improve the electrolarynx include the use of Mel frequency cepstral coefficients (MFCC) to extract essential features from electrolarynx output and attenuate generic vibrational noise as well as expanding the frequency range of the traditional electrolarynx (T. Masaki, personal communication, October 20, 2021; Padmini, 2017). This project expands on the research above by creating an ANN that utilizes both video processing and audio processing to predict speech as accurately as possible.

Proposed Methods

Preparation

Five English-speaking adults, both male and female, were trained to read *The Rainbow Passage*, a short passage containing all the phonemes in the English language. Videos of their lips while reading the passage with and without an electrolarynx were recorded. This process was repeated for a total of 20 times for each participant for a total of 200 videos. Video and audio output will be extracted from these videos and will be used as input for two different pipelines in the final algorithm: video processing and audio processing.

Pipelines

Feature extraction for the video processing pipeline will be conducted using DeepLabCut, a software intended to capture the geometrical configuration of user-selected body parts using deep learning (Mathis et al., 2018). DeepLabCut will utilize a user-selected ANN to determine four selected points on the lip for video frames, which will be used as input for a CNN developed in PyTorch. This CNN will output predicted speech based on the positions of the lips over time.

Feature extraction for the audio processing pipeline will be conducted using MFCC, an audio feature extraction technique that emphasizes frequencies common to human speech while deemphasizing others (Dave, 2013). The output from this process will be used as an input to an ANN that includes both a CNN and a long short-term memory (LSTM) network that will predict speech based both on current and previous electrolarynx output.

Final Steps

After training the combined pipelines on the Rainbow Passage videos, the model will be refined using short clips of conversational speech. The model will be tested on videos of laryngectomees reading another passage similar to the Rainbow Passage as well as laryngectomees speaking conversationally (Figure 1). The final algorithm will be refined until it predicts speech using lip position and respective electrolarynx output with an accuracy of at least 85%, with the ideal accuracy being as close to 100% as possible.

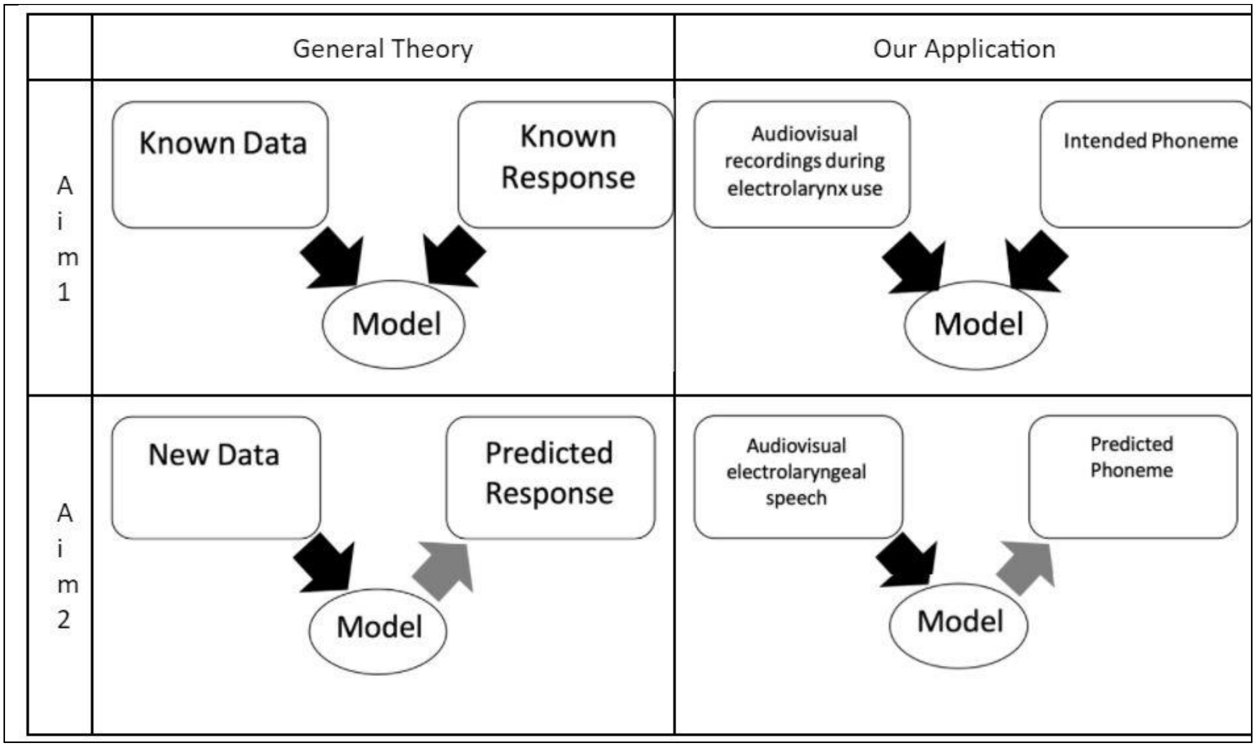

Figure 1. Basic ANN model for training (top row) and testing (bottom row) data (Jonas, 2021).

Significance

If successful, this algorithm will be able to mitigate several issues common to laryngectomees, including the inability to speak in noisy environments, necessity for multiple speech therapy appointments, and the general decrease in quality of life. The use of computer-generated speech will allow laryngectomees to change the volume of speech whenever necessary, allowing clear speech in noisy environments.

The use of only five subjects for training data and the focus on the English language are drawbacks to this algorithm. All subjects were trained by a speech pathologist and have no accent, so the algorithm may have difficulty predicting conversational speech, especially if the laryngectomee has an accent. Additionally, other languages utilize different lip movements and have different phonemes than the English language, which may result in difficulty applying the algorithm to other languages.

Responsible Innovation: A Framework for Bias in Machine Learning

Many of the drawbacks previously listed are potential sources of bias in the algorithm, an important topic that merits discussion. The possibility of bias in machine learning algorithms has become progressively more known in recent years, particularly with prominent social issues such as the development of the COMPAS sentencing algorithm. Many types of bias can infiltrate a machine learning algorithm, usually as a result of poor feature selection, bad training data or lack of transparency in the algorithm creation process. The COMPAS sentencing algorithm was marked by racial bias: prisoners of color were more likely to be deemed dangerous, while white prisoners who eventually re-offended were marked as low risk. This was eventually found to be a result of using an inmate's neighborhood and family crime history as factors in the decision (Martin, 2019; Mehrabi et al., 2021; Yates et al., 2013). Gender bias is frequently found in natural language processing as a result of the text the algorithms train on (Leavy, 2018). Due to an overload of American pictures in the training data, an object recognition algorithm recognized a Western wedding but not an Indian wedding (de Vries, Misra, Wang, & van der Maaten, 2019). Finally, lack of transparency in algorithm creation can lead to user ignorance, which increases the likelihood of bias (Martin, 2019).

When bias is found in machine learning algorithms in the health field, the consequences can be deadly. For example, an algorithm that predicted health outcomes assigned the same level of risk to patients of color who were much sicker than their white counterparts (Obermeyer, Powers, Vogeli, & Mullainathan, 2019). Bias in medical algorithms is exacerbated by the explicit and implicit biases present in healthcare professionals and therefore data collection (Starke, De Clercq, & Elger, 2021). Since this project seeks to improve both communication and the quality of life of laryngectomees, it is of utmost importance that the algorithm allows people of different races, ethnicities, genders, socioeconomic statuses, and disability statuses to communicate equitably.

Since many biased algorithms begin with a biased dataset, the elimination of bias in a machine learning algorithm starts with data collection. Since biases may be implicit and therefore unknown to the creator of an algorithm, bias mitigation may seem like a daunting task. Therefore, a sociotechnical framework might inform developers in their search for a fair data set and algorithm.

The Responsible Innovation framework, developed by Stilgoe et al. (2013), seeks to give a detailed plan for developing a technology with as few negative consequences as possible. Stilgoe first acknowledges that most innovators seek to develop their products responsibly without knowingly inserting bias into their algorithms. However, technology development is plagued by unintended consequences which need to be recognized before they happen. In order to recognize and prevent unintended consequences such as bias, developers should consider a framework with four critical dimensions: anticipation, reflectiveness, deliberation, and responsiveness (Stilgoe, Owen, & Macnaghten, 2013).

Anticipation asks the question “what if?” It seeks to determine the possible unintended consequences by thoroughly examining the technology. Reflectiveness is an extension of anticipation and seeks to discuss what is known, such as the purposes and motivations of the technology, as well as what is unknown, such as areas of ignorance and questions. Deliberation entails the opening of this conversation about possible biases to stakeholders. This allows for a collaborative environment that acknowledges a variety of perspectives. Responsiveness is a secondary, iterative measure that reflects on past progress in order to direct the future progression of the technology (Stilgoe et al., 2013). With respect to bias, anticipation and reflection entail the examination of the dataset, methods of data collection, and the algorithm itself for potential sources of bias, with an emphasis on analyzing ways the algorithm may behave unexpectedly. Deliberation with stakeholders may open developers up to their own implicit biases. Finally, biases in a machine learning algorithm should be investigated after feature extraction, training, validation, and testing as a form of responsiveness.

When moving from the theoretical to the practical, the responsible innovation framework offers several concrete examples of both proactive and irresponsible behavior during the development of a technology such as a machine learning algorithm. For example, the ethics of an algorithm should be a design factor rather than a constraint (von Schomberg, 2013). The goal should move from making an algorithm that won’t cause any trouble to making an algorithm that serves all targeted stakeholders equitably. In contrast, the tendency to produce technology as quickly as possible due to demands from stakeholders can easily result in unintended consequences (von Schomberg, 2013). Further research will be conducted on specific action items necessary for bias mitigation within machine learning algorithms, specifically within the context of the project.

Research Design

The responsible innovation framework leads me to ask: How can sources of bias be mitigated in machine learning algorithms, especially algorithms that involve the healthcare sector? This question will be answered through a comprehensive literature review followed by interviews. Multiple journal articles that describe a machine learning algorithm and at least one bias mitigation technique used will be reviewed. These mitigation techniques will then be analyzed within the context of the responsible innovation framework. A further literature review of algorithms that failed due to bias will be conducted in order to determine which dimension of the framework is most often ignored.

Next, interviews with stakeholders in the algorithm development process will be conducted. The interviews will address the dangers of bias in machine learning, current mitigation techniques, and application of the responsible innovation framework to a project related to the stakeholder. The first group of interviewees will be algorithm developers who can provide specific insight on bias mitigation measures that are currently used. I will contact professors from the CS department who work in machine learning to find this segment of interviewees. Doctors will comprise the next segment of interviewees and can provide insight on the effects of bias in a machine learning algorithm on their patients. This may prove more difficult to find willing interviewees, but my medical advisor will be interviewed in the worst-case scenario. The final segment of interviewees will be anyone knowledgeable about machine learning who will play the role of a patient. These patients can voice concerns about algorithm transparency and bias against any marginalized group they may be a part of. I will use a Reddit survey under a relevant machine learning thread in order to gather information from this segment. As backup, I will identify fellow students who took the machine learning class that sparked my interest in machine learning. Overall, the interviews will provide a comprehensive view of bias in machine learning as well as the current state of mitigation techniques in algorithm development.

In a final step, the results from the literature review and interviews will be synthesized into a final report that details the current state of bias mitigation in terms of the responsible innovation framework. An analysis of the elements of the framework that are utilized least as well as irresponsible behaviors that are most pervasive will clarify the responsible algorithm development procedure detailed in the responsible innovation framework.

Conclusion

This project seeks to improve post-laryngectomy communication by developing an ANN that uses lip movements and the traditionally used electrolarynx output to predict speech. The use of an ANN necessitates a discussion about bias in machine learning algorithms and methods for mitigating bias, particularly in a healthcare setting. The responsible innovation framework was used to frame a plan for bias mitigation emphasizing anticipation, reflectiveness, deliberation, and responsiveness.

It is expected that most stakeholders would be concerned about bias, but active mitigation techniques would be less frequent. Common irresponsible behaviors will almost certainly include a “technology push” to produce technology as quickly as possible due to deadlines as well as a lack of algorithm transparency. Informed by the results of this research, the developers of this project will be able to identify sources of bias and create an algorithm that not only improves communication but also treats all laryngectomees equitably.

References

Antin, F., Breheret, R., Goineau, A., Capitain, O., & Laccourreye, L. (2021). Rehabilitation following total laryngectomy: Oncologic, functional, socio-occupational and psychological aspects. *European Annals of Otorhinolaryngology, Head and Neck Diseases*, *138*(1), 19–22. https://doi.org/10.1016/j.anorl.2020.06.006

Assael, Y. M., Shillingford, B., Whiteson, S., & de Freitas, N. (2016). LipNet: End-to-End Sentence-level Lipreading. *ArXiv:1611.01599 [Cs]*. Retrieved from http://arxiv.org/abs/1611.01599

Carr, M. M., Schmidbauer, J. A., Majaess, L., & Smith, R. L. (2000). Communication after laryngectomy: An assessment of quality of life. *Otolaryngology–Head and Neck Surgery*, *122*(1), 39–43. https://doi.org/10.1016/S0194-5998(00)70141-0

Cox, S. R., & Doyle, P. C. (2014). The Influence of Electrolarynx Use on Postlaryngectomy Voice-Related Quality of Life. *Otolaryngology–Head and Neck Surgery*, *150*(6), 1005–1009. https://doi.org/10.1177/0194599814524704

Dave, N. (2013). Feature extraction methods LPC, PLP and MFCC in speech recognition. *International Journal For Advance Research in Engineering And Technology(ISSN 2320-6802), Volume 1*.

de Vries, T., Misra, I., Wang, C., & van der Maaten, L. (2019). *Does Object Recognition Work for Everyone?* 52–59. Retrieved from https://openaccess.thecvf.com/content_CVPRW_2019/html/cv4gc/de_Vries_Does_Object_Recognition_Work_for_Everyone_CVPRW_2019_paper.html

Gates, G. A., Ryan, W., Cooper, J. C., Lawlis, G. F., Cantu, E., Hayashi, T., … Hearne, E. (1982). Current status of laryngectomee rehabilitation: I. Results of therapy. *American Journal of Otolaryngology*, *3*(1), 1–7. https://doi.org/10.1016/s0196-0709(82)80025-2

Kohlberg, G. D., Gal, Y. (Kobi), & Lalwani, A. K. (2016). Development of a Low-Cost, Noninvasive, Portable Visual Speech Recognition Program. *Annals of Otology, Rhinology & Laryngology*, *125*(9), 752–757. https://doi.org/10.1177/0003489416650689

Leavy, S. (2018). Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning. *Proceedings of the 1st International Workshop on Gender Equality in Software Engineering*, 14–16. Gothenburg Sweden: ACM. https://doi.org/10.1145/3195570.3195580

Martin, K. (2019). Ethical Implications and Accountability of Algorithms. *Journal of Business Ethics*, *160*(4), 835–850. https://doi.org/10.1007/s10551-018-3921-3

Mathis, A., Mamidanna, P., Cury, K. M., Abe, T., Murthy, V. N., Mathis, M. W., & Bethge, M. (2018). DeepLabCut: Markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, *21*(9), 1281–1289. https://doi.org/10.1038/s41593-018-0209-y

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, *54*(6), 1–35. https://doi.org/10.1145/3457607

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

O'Mahony, N., Campbell, S., Carvalho, A., Harapanahalli, S., Hernandez, G. V., Krpalkova, L., … Walsh, J. (2020). Deep Learning vs. Traditional Computer Vision. In K. Arai & S. Kapoor (Eds.), *Advances in Computer Vision* (pp. 128–144). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-17795-9_10

Padmini, L. (2017). Neural Network based New Bionic Electro Larynx Speech System. *International Journal of Engineering Research*, *5*(13), 5.

Starke, G., De Clercq, E., & Elger, B. S. (2021). Towards a pragmatist dealing with algorithmic bias in medical machine learning. *Medicine, Health Care and Philosophy*, *24*(3), 341–349. https://doi.org/10.1007/s11019-021-10008-5

Stilgoe, J., Owen, R., & Macnaghten, P. (2013). Developing a framework for responsible innovation. *Research Policy*, *42*(9), 1568–1580. https://doi.org/10.1016/j.respol.2013.05.008

von Schomberg, R. (2013). A Vision of Responsible Research and Innovation. In R. Owen, J. Bessant, & M. Heintz (Eds.), *Responsible Innovation* (pp. 51–74). Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118551424.ch3

Wulff, N. B., Højager, A., Wessel, I., Dalton, S. O., & Homøe, P. (2021). Health-Related Quality of Life Following Total Laryngectomy: A Systematic Review. *The Laryngoscope, 131*(4), 820–831. https://doi.org/10.1002/lary.29027

Yates, D. J., Gulati, G. J. J., & Weiss, J. W. (2013). Understanding the Impact of Policy, Regulation and Governance on Mobile Broadband Diffusion. *2013 46th Hawaii International Conference on System Sciences*, 2852–2861. Wailea, HI, USA: IEEE. https://doi.org/10.1109/HICSS.2013.583