IMPROVING ELECTROLARYNX OUTPUT USING MACHINE LEARNING ON ACOUSTIC AND VISUAL

RELATIONSHIP BETWEEN LANGUAGES AND APPLICABILITY TO ALGORITHM DESIGN

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Biomedical Engineering

By
Medhini Rachamallu

November 1, 2021

Technical Team Members:
Sameer Agrawal
Surabhi Ghatti
Katherine Taylor

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.

ADVISORS

Rider Foley, Department of Engineering and Society

Haibo Dong, Department of Aerospace and Mechanical Engineering

Introduction

Every year, over 3000 patients undergo a total laryngectomy, full removal of the voice box, to treat laryngeal cancer (Kohlberg, Gal, & Lalwani, 2016). During the procedure, the larynx is removed, and the trachea (the windpipe) is separated from the throat, severing the connection between the lungs and the mouth (Figure 1). This prevents movement of air to the mouth which creates sound (American Cancer Society [ACS], 2021). Therefore, laryngectomees, patients who have undergone a laryngectomy, lose their ability to speak and are forced to acquire alternative means of communication including gesture, writing, and voice restoration therapies.

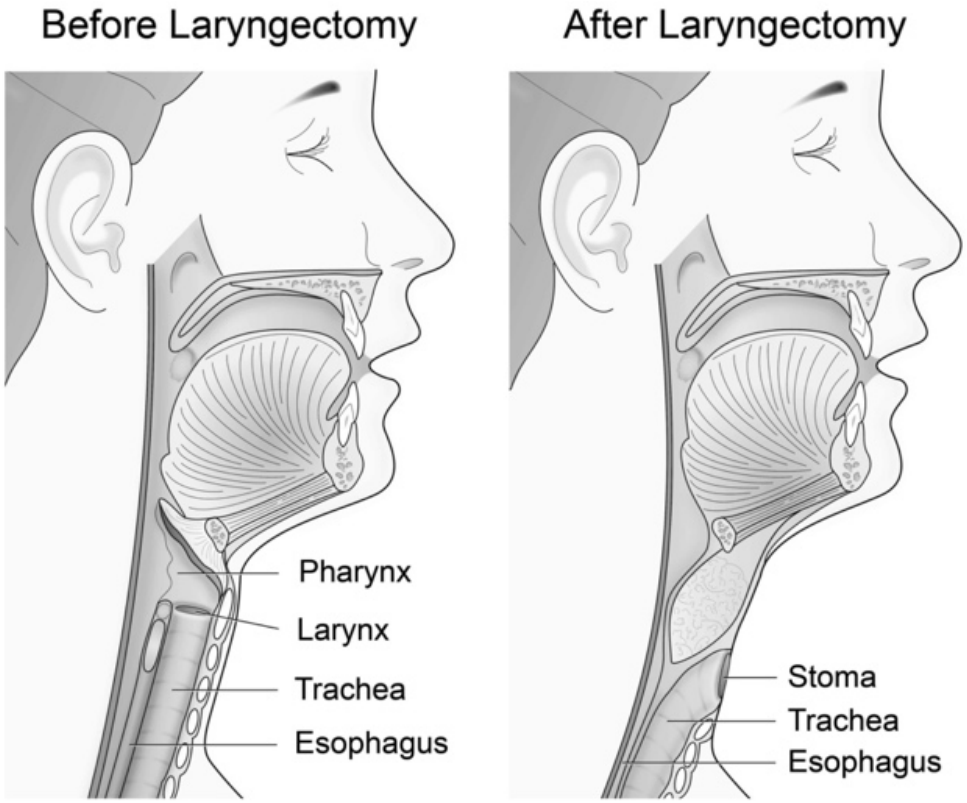

Figure 1. *Physical Changes Due to Total Laryngectomy*. This figure demonstrates how the trachea changes after a laryngectomy (Albirmawy, Elsheikh, Saafan, & Elsheikh, 2006). The air cannot flow from the lungs to the mouth anymore, which means no sounds can be created.

The most common voice restoration therapy is trachea-esophageal puncture (TEP), which is a surgical procedure that allows patients to redirect air out of their mouth (ACS, 2021). Roughly 25% of patients face complications following TEP (Andrews, Mickel, Monahan, Hanson, & Ward, 1987), including recurring infections, pneumonia, and sometimes death (Albirmawy et al., 2006). These complications can arise soon after surgery or many years later (Wang et al., 1991), reducing predictability.

The electrolarynx, a battery-operated device that creates a mechanical voice, is another voice restoration method, but it is non-invasive. Less complications are associated with the electrolarynx, but it also produces reduced intelligible speech when compared with TEP. Although it is a very common form of voice restoration, traditional electrolarynxes produce little, if any, pitch variation, creating a single fundamental voice frequency (F0) (Liana Guo, 2016). Regular speech contains F0 contours, which humans naturally interpret as tone in speech (Eady, 1982). Additionally, all widely used electrolarynx models require users to operate the device with one hand and find the correct location to press the vibration, which reduces accessibility since one hand is occupied. Still, electrolarynx technology is vital due to its non-invasive nature and because it serves as the main backup for those who face complications in other voice restoration therapies (ACS, 2021).

The biggest pitfall of a total laryngectomy is the subsequential depreciation in quality of life that patients experience. Along with changes to breathing and swallowing, these patients can no longer communicate the way that they used to, leading to isolation, mental health issues, and a decrease in quality of healthcare (Souza et al., 2020). Improving electrolarynx technology provides laryngectomees with a non-invasive, simple way to be able to speak again, which could drastically improve their quality of life. Therefore, my technical project aims to create and train

an Artificial Neural Network (ANN), which consists of a Convolutional Neural Net (ConvNet) and a Long Short-Term Memory (LSTM) network, using visual and auditory signals from videos of laryngectomee's and non-laryngectomees with and without an electrolarynx to create a computer generated, intelligible, "normal" voice. However, since all our data is from patients speaking in the English language, this project cannot help patients around the world who speak other languages. I will explore how languages are related to understand how we can incorporate these relationships while designing algorithms.

Technical Topic

Although the electrolarynx is associated with far less complications than TEP, it produces a more robotic, mechanical voice that can be unintelligible. Dissatisfaction with a laryngectomee's ability to communicate can lead to a depreciation in quality of life both with families and healthcare providers, an increase in isolation, and the development of mental health issues (Souza et al., 2020). Therefore, improving electrolarynx technology provides laryngectomees with a non-invasive, simple way to be able to speak again through the help of hospital-provided speech therapists.

My project aims to create and train an ANN, which consists of a ConvNet and a LSTM network, using visual and auditory signals from an electrolarynx to create a computer generated, intelligible, "normal" voice to increase speech intelligibility. Processing auditory data using Mel frequency cepstral coefficients (MFCC) (Qian, Wang, Zhang, Liu, & Niu, 2019) for feature extraction and training using ConvNets has shown reduced error rates by 6-10% (S, U, K, & Padmini, 2018). Combined with LSTMs, studies have shown that word recognition can have an error rate as low as 6% with only auditory data (W. Wang, Yang, & Yang, 2020). This project

aims to combine auditory data with visual data to improve accuracy overall and in settings where only one of the two might be available.

New electrolarynxes have been created which attempt to modulate F0 frequency to convey tone. These models use varying amounts of finger pressure on a single button, control expiration pressure from the neck, filter electromyographic signals from neck muscle contractions, and adjust forearm tilt movement to create changes in F0 frequency, but they convey natural intonation patterns to only varying degrees of success (Liana Guo, 2016). Additionally, two electrolarynx models currently being developed in Japan models also remove the use of a hand: the first consists of two frequency modulators that attach around the neck as a band and the second is a retainer-style frequency modulator that fits into the mouth (Takeuchi, Masaki, 2021). Due to the stabilization of the vibrator, these models also produce more intelligible sound output.

Previously, Microsoft's Xbox 360 Kinect Model 1414 was used to track a subject's lip movements utilizing three locations to calculate distances, applying facial recognition software, and training an ANN with this data. Optimization for this network produced ~77% accuracy in the ability to correctly identify sentences based on lip movements (Kohlberg et al., 2016). A similar model can be applied to the electrolarynx to track articulatory movements and improve intelligible speech production. This study corroborates that lip reading data can be trained using ANNs. Other studies have also used ConvNets and LSTMs to train lip reading datasets (Fernandez-Lopez & Sukno, 2018).

The first part of our technical project involves designing an ANN to map English phonemes to visual and acoustic electrolarynx in individuals with functioning larynges using *The Rainbow Passage*, a well-known passage containing all the phonemes in the English language.

Five English-speaking adults were recruited, trained to read this passage by a speech therapist, and recorded with and without an electrolarynx. After the ANN has been trained, it will be refined using conversational speech between these adults. The second part of this project involves applying the trained ANN to conversational speech in laryngectomees. Five English-speaking laryngectomee adults will be recruited to read *The Caterpillar Passage*, a passage with similar elements, and have conversational speech using the electrolarynx.

Both auditory and visual data have their own drawbacks. With the auditory data, the electrolarynx produces a very high background noise that needs to be removed before any features can be extracted using MFCC and Fourier Transforms. Additionally, the presence of accents even in the English language could create some obstacles, especially if English is not the speaker's native language (Kavanagh, 2007). Currently, the data collected for this project includes patients and volunteers who have been trained on how to speak with electrolarynx, but since accents do carry through the electrolarynx, running more conversational data through this model could pose potential problems. With visual lip-reading data, there are many phonemes that have the same mouth positioning. The phonemes *p* and *b* are nearly indistinguishable through video because the change in pronunciation occurs inside the mouth. The same happens for *k* and *g* since only the tongue position inside the mouth changes (Fernandez-Lopez & Sukno, 2018). However, combining both audio and visual data could be powerful if the datasets are able to complement each other.

Although the audio and visual data might provide a high accuracy together, one major drawback of this project is that it focuses solely on the English language. Other areas around the world, such as South America and Africa, have a large prevalence of total laryngectomies (Fagan, Lentin, & Quail, 2013; Vartanian, Carrara-de-Angelis, & Kowalski, 2013), but their

languages differ significantly from the English language. In fact, some sounds in other languages do not even exist in the English language, meaning that they would not be captured by the two passages used to train the ANN.

STS Topic

Although the data for our technical project consists of English speakers trained by a speech therapist, patients internationally use an electrolarynx to communicate after a laryngectomy. I will seek to understand the relationship between languages and how those relationships can help develop algorithms.

Generally, languages can be grouped into two broad categories: tonal and stress languages (Eady, 1982). Stress languages include those such as English, German, etc. (Cahill & Tiberius, 2004; Dalsgaard, Andersen, Hesselager, & Petek, 1996), where the syllables of the word are characterized by varying degrees of emphasis. This stress causes an increase in F0 frequency and duration of the symbol, creating a sentence melody (Eady, 1982). Tonal languages include those such as Mandarin, Thai, etc. (Liana Guo, 2016), where intonation changes within a single word can differentiate between meanings. The overall sentence intonation in Mandarin can be explained as small ripples on a larger wave (Eady, 1982).

To compare stress and tonal languages, I will use the Linguistic Framework, consisting of three parts: 1) language has a structure, 2) when attempting to investigate the structure, a language needs to be in place, and 3) making conventions is part of the investigation (Torfehnezhad, 2016). Carnap's use of this structure defines pragmatics, semantics, and syntax as the most important part of a language system, where language can be analyzed from pragmatics to syntax or the opposite direction. Pragmatics focuses most on the relationship

between the speaker and the world while the other two parts are focused on solely the speaker. Pragmatics is defined by how the speakers of a language create signs for objects, etc. to communicate in their community, understand events, and construct theories about the world. Semantics is defined as the actual words assigned to these objects and the rules for speaking. Finally, syntax provides the interpretation of a language by understanding the expression present (Torfehnezhad, 2016). My analysis will focus mostly on semantics and syntax (Chomsky, 1955) because language datasets used in algorithms incorporate words and phonemes, which change between languages.

Semantics and syntax influence intonational boundaries in both stress and tonal languages to convey side remarks, clauses, vocatives, and more. The presence of a pause can indicate that upcoming information is new or important (Watson & Gibson, 2004). There are three major theories governing the placement of intonational boundaries. The Cooper and Paccia-Cooper theory hypothesizes that more syntactic constituents beginning or ending a phrase indicate a greater likelihood of an intonational boundary. The Gee and Grosjean theory proposes an eight-step approach to predict boundaries which includes finding phonological phrases and the positioning of those phrases with respect to verbs. The third theory, Ferreira's, examines both syntactic and semantic structure to identify intonational phrasing (Watson & Gibson, 2004). Together, these theories can be leveraged to isolate the use and importance of tone across stress and tonal languages.

A major part of semantics includes phonemes which is the minimal contrastive sound unit of a language. Changing a phoneme could change the meaning of a word such as in *bit* and *pit* (Kavanagh, 2007). Allophones are different realizations of phonemes such as the sounds that *k* can create. To be an allophone, the sound distribution must be predictable and if one phoneme

is exchanged for the other in the same context, the substitution must not lead to a difference in meaning. Many other semantic structures exist such as dipthongs, the liquid R and L, and fricatives (Kavanagh, 2007). These constructions vary between languages even when they are of the same type. For example, Japanese and English are both stress languages, but Japanese contains 17 consonants and 5 vowels while English contains 24 constants and 20 vowels as phonemes (Kavanagh, 2007). When learning English as a second language, Japanese speakers carry an accent because English contains different phonetic constructions than Japanese. The phonological sounds in one's native language can influence the production of sound in the second language (Kavanagh, 2007). These subtleties increase the complexity of understanding the relationship between languages.

Language similarities can be characterized using metaphonemes, which are closely related phonemes found in multiple languages (Cahill & Tiberius, 2004). Metaphonemes are like allophones, but interchanging metaphonemes does not alter the meaning of a word. One example is the English word *cat*, the Dutch word *kat*, and the German word *katze* (Cahill & Tiberius, 2004). The vowels in these three words are slightly different even though the consonants are identical, creating slight accents during pronunciation. If an English speaker were to hear any of these three words, they would be able to understand the meaning and recognize the presence of an accent. This concept has evolved for some time now from keysymbols and archiphonemes as linguists attempt to construct a universal set of phonemes (Cahill & Tiberius, 2004). Creating a universal set of phonemes could be vital to developing algorithms that cross language-imposed boundaries.

Research Question and Methods

Using the Linguistic Framework, I will be comparing tonal and stress languages while focusing on semantics and syntax. My technical project involves training an ANN with English audio and visual data to improve speech intelligibility from the electrolarynx, which can be beneficial for patients that speak other languages as well. Therefore, I seek to understand the semantics and syntax that remain the same across languages and whether a universal set of phonemes can be created to optimize datasets and algorithms.

One venue for research will be through in-person interviews. I will begin by contacting the speech therapist that was involved in filming the videos for our dataset and continue by speaking with laryngectomees. My primary goal will be interviewing bilingual patients who speak English and use an electrolarynx because it will provide first-hand accounts on their struggles and what they think are the similarities and differences between the languages they speak when using the electrolarynx.

I will apply the Linguistic Framework to extract phonemes and uses of tone in tonal and stress languages from literature. The tonal languages used will be Mandarin and Thai while the stress languages will be English, German, and Japanese. These languages have been chosen because there is a lot of research that has been done on them both individually and in comparison with English. I will then perform a comparative analysis to determine whether metaphonemes exist between these languages to understand if a universal set of phonemes can be created between the two major language classes. The information from the interviews will also be applied at this stage to determine whether the uses of intonation in each language can be conveyed through the electrolarynx and whether they are necessary for adequate speech comprehension.

Conclusion

Partial and total laryngectomies are the most common treatment for laryngeal cancer (ACS) 2021), but they lead to a large depreciation in quality of life due to speech loss (Souza et al., 2020). TEP and electrolarynxes are the main voice restoration methods. TEP is associated with many complications unlike electrolarynxes, but TEP more intelligible speech. Therefore, improving electrolarynx technology to improve intelligibility can improve choices for patients.

My technical project focuses on creating and training an ANN, which consists of a ConvNet and a LSTM network, using visual and auditory signals from an electrolarynx to create a computer generated, intelligible, "normal" voice to increase speech intelligibility. All our data is in English, but patients around the world receive laryngectomies. Therefore, my STS topic focuses on the relationships between languages and how that can help design algorithms. Using the Linguistic Framework, I will analyze the semantics and syntax to investigate whether a universal set of phonemes can be created with stress and tonal languages.

References

Albirmawy, O. A., Elsheikh, M. N., Saafan, M. E., & Elsheikh, E. (2006). Managing problems with tracheoesophageal puncture for alaryngeal voice rehabilitation. *The Journal of Laryngology & Otology*, *120*(6), 470–477. https://doi.org/10.1017/S0022215106000752

American Cancer Society (ACS). (2021). Throat Cancer Follow-Up | Living as a Throat Cancer Survivor. Retrieved September 26, 2021, from https://www.cancer.org/cancer/laryngeal-and-hypopharyngeal-cancer/after-treatment/follow-up.html

Andrews, J. C., Mickel, R. A., Monahan, G. P., Hanson, D. G., & Ward, P. H. (1987). Major complications following tracheoesophageal puncture for voice rehabilitation. *The Laryngoscope*, *97*(5), 562–567. https://doi.org/10.1288/00005537-198705000-00005

Cahill, L., & Tiberius, C. (2004). *Cross-Linguistic Phoneme Correspondences*. https://doi.org/10.3115/1071884.1071906

Chomsky, N. (1955). Logical Syntax and Semantics: Their Linguistic Relevance. *Language*, *31*(1), 36–45. https://doi.org/10.2307/410891

Dalsgaard, P., Andersen, O., Hesselager, H., & Petek, B. (1996). Language-identification using language-dependent phonemes and language-independent speech units. *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, *3*, 1808–1811 vol.3. https://doi.org/10.1109/ICSLP.1996.607981

Eady, S. J. (1982). Differences in the F0 Patterns of Speech: Tone Language Versus Stress Language. *Language and Speech*, *25*(1), 29–42. https://doi.org/10.1177/002383098202500103

Fagan, J. J., Lentin, R., & Quail, G. (2013). International practice of laryngectomy rehabilitation interventions: A perspective from South Africa. *Current Opinion in Otolaryngology &*

Head and Neck Surgery, *21*(3), 199–204. https://doi.org/10.1097/MOO.0b013e328360c3c1

Fernandez-Lopez, A., & Sukno, F. M. (2018). Survey on automatic lip-reading in the era of deep learning. *Image and Vision Computing*, *78*, 53–72. https://doi.org/10.1016/j.imavis.2018.07.002

Kavanagh, B. (2007). *The phonemes of Japanese and English: A contrastive analysis study*. https://doi.org/10.24552/00001866

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

Liana Guo, K. N. (2016). Generating Tonal Distinctions in Mandarin Chinese Using an Electrolarynx with Preprogrammed Tone Patterns. *Speech Communication*, *78*, 34. https://doi.org/10.1016/j.specom.2016.01.002

Qian, Z., Wang, L., Zhang, S., Liu, C., & Niu, H. (2019). Mandarin Electrolaryngeal Speech Recognition Based on WaveNet-CTC. *Journal of Speech, Language, and Hearing Research*, *62*(7), 2203–2212. https://doi.org/10.1044/2019_JSLHR-S-18-0313

S, A., U, K., K, K. V., & Padmini, M. L. (2018). Neural Network based New Bionic Electro Larynx Speech System. *International Journal of Engineering Research & Technology*, *5*(13). Retrieved from https://www.ijert.org/research/neural-network-based-new-bionic-electro-larynx-speech-system-IJERTCONV5IS13154.pdf, https://www.ijert.org/neural-network-based-new-bionic-electro-larynx-speech-system

Souza, F. G. R., Santos, I. C., Bergmann, A., Thuler, L. C. S., Freitas, A. S., Freitas, E. Q., & Dias, F. L. (2020). Quality of life after total laryngectomy: Impact of different vocal

rehabilitation methods in a middle income country. *Health and Quality of Life Outcomes*, *18*(1), 92. https://doi.org/10.1186/s12955-020-1281-z

Takeuchi, Masaki. (2021, October 20). *Information about Syrinx* [Video Conference].

Torfehnezhad, P. (2016). *In Carnap's Defense: A survey on the concept of a linguistic framework in Carnap's philosophy*. *9*(1), 3–30.

Vartanian, J. G., Carrara-de-Angelis, E., & Kowalski, L. P. (2013). Practice of laryngectomy rehabilitation interventions: A perspective from South America. *Current Opinion in Otolaryngology & Head and Neck Surgery*, *21*(3), 212–217. https://doi.org/10.1097/MOO.0b013e328361067b

Wang, R. C., Bui, T., Sauris, E., Ditkoff, M., Anand, V., & Klatsky, I. A. (1991). Long-term Problems in Patients With Tracheoesophageal Puncture. *Archives of Otolaryngology–Head & Neck Surgery*, *117*(11), 1273–1276. https://doi.org/10.1001/archotol.1991.01870230089014

Wang, W., Yang, X., & Yang, H. (2020). End-to-End Low-Resource Speech Recognition with a Deep CNN-LSTM Encoder. *2020 IEEE 3rd International Conference on Information Communication and Signal Processing (ICICSP)*, 158–162. https://doi.org/10.1109/ICICSP50920.2020.9232119

Watson, D., & Gibson, E. (2004). The relationship between intonational phrasing and syntactic structure in language production. *Language and Cognitive Processes*, *19*(6), 713–755. https://doi.org/10.1080/01690960444000070