

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

A Technical Report submitted to the Department of Biomedical Engineering

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia


In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

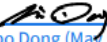
Surabhi Ghatti
Spring Semester 2022

Capstone Project Team Members
Sameer Agrawal
Medhini Rachamalla
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

Signature Surabhi Date 05/07/2022
Surabhi Ghatti

Approved  Date 05/05/2022
James J. Daniero (May 5, 2022 12:15 EDT)
James J. Daniero, M.D., Department of Otolaryngology - Head and Neck Surgery

Approved  Date 05/05/2022
Haibo Dong (May 5, 2022 09:49 EDT)
Haibo Dong, Ph.D., Department of Mechanical and Aerospace Engineering

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

Sameer Agrawal^a, Surabhi Ghatti^b, Medhini Rachamalla^c, Katherine Taylor^d

^a sa9nc@virginia.edu

^b sg2dk@virginia.edu

^c mr3dn@virginia.edu

^d kmt2ns@virginia.edu

Abstract

After patients undergo a total laryngectomy, alternate methods of speech are necessary to facilitate normal communication. One non-invasive voice restoration method is the electrolarynx, which has fewer complications than other forms of voice restoration therapy but results in a lower quality of speech. The proposed solution to this dilemma is the creation of a dual-pipeline deep neural network (DNN) to translate lip movements and audio output from electrolarynx use into intelligible speech. Videos of study participants reciting the *Rainbow Passage*, a phonetically balanced passage, as well as speaking conversationally for sixty seconds were recorded with and without an electrolarynx. The audio pipeline iterated through a number of ways to extract the necessary audio data from each video, including Mel Frequency Cepstral Coefficients (MFCCs). A binary classification DNN was used to determine if the MFCCs were a good fit for distinguishing between electrolarynx and non-electrolarynx data. Despite a testing accuracy of 61%, it was determined that the error was not decreasing as iterations decreased, showing a lack of learning on the part of the DNN. The video pipeline utilized a software called DeepLabCut to predict points on the lip for both *Rainbow Passage* and conversational speech videos using a convolutional neural network (CNN) trained on frames labeled with four points on the lip. DeepLabCut effectively predicted these four points for the conversational speech videos, with a mean pixel error of 5.11 pixels. In summary, the audio pipeline successfully determined that MFCCs are not a viable audio analysis technique for this project, while the video pipeline was able to successfully predict four points on a lip for speakers with similar distances from the camera. The next steps in this project include researching other methods of audio analysis to extract essential features for the DNN, labeling the frames with the associated phoneme for the next steps in the video pipeline, and shortening the videos to facilitate this labeling process.

Keywords: electrolarynx, voice restoration therapies, DNN, MFCCs, DeepLabCut, neural network

Introduction

Clinical Background

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 cancer¹. During this procedure, the larynx is removed, and the trachea is separated from the throat, severing the connection between the lungs and the mouth. As a result, laryngectomees lose their ability to speak and are forced to acquire alternative means of communication. These possibilities include gesturing, writing, and voice restoration therapies. 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^{2,3}. The gold standard for voice restoration is the tracheoesophageal puncture (TEP), a surgical procedure that allows patients to redirect air out of their mouth. Despite its designation as the gold standard, roughly 25% of TEP recipients face

complications, including recurring infections, pneumonia, and sometimes death^{4,5}. The onset of complications can range from soon after surgery to years later, reducing predictability⁶.

An alternative to TEP is the electrolarynx, a battery-operated device placed under the chin that emits vibrations that are transmitted through the skin to the throat. These vibrations are shaped into words using the lips, tongue, and teeth, creating a mechanical voice. The electrolarynx is preferred by some to other forms of voice restoration because it is non-invasive, has no complications, and is most cost-effective⁷. Furthermore, it serves as an easy backup for those who face complications from other voice restoration therapies⁷. Similar to TEP, the electrolarynx comes with significant drawbacks. While it is considered the easiest and fastest voice restoration technique to learn, speaking with an electrolarynx effectively still takes significant time and effort from both the patient and a speech therapist, which taxes the healthcare system⁷.

Furthermore, the voice produced by the electrolarynx is mechanical, monotonic, and difficult to understand. This lack of intelligibility is exacerbated by the presence of generic vibrational noise that drowns out the voice of the speaker^{8,9}. Additionally, the current electrolarynx requires users to operate the device with one hand and place the electrolarynx on the correct location on the throat, reducing accessibility¹⁰.

Regardless of the voice restoration method used, laryngectomees experience a reduced quality of life for multiple reasons. Physically, laryngectomees struggle with difficulty swallowing and may be subject to postoperative chemotherapy if there is a recurrence of cancer. Emotionally, laryngectomees are taxed with the burden of permanently losing the ability to speak normally as well as the possibility of cancer recurrence^{9,11}. Financially, laryngectomees must attend multiple sessions with a speech therapist in order to learn how to use the chosen voice restoration therapy, which can pose a considerable burden to those who are underinsured. Environmentally, it is difficult to communicate effectively in noisy settings due to the presence of generic vibrational noise and limited volume capacity; laryngectomees may also experience a reduction in quality of healthcare due to the inability to communicate clearly^{7,12,13}. The burden placed on laryngectomees has led to social isolation and reports of fewer friends, increased mental health issues such as anxiety and depression, and decreased self-estimated quality of life scores among the population^{7,10,12,13}.

Technical Background

In order to improve the quality of life of a laryngectomee to the highest extent possible, an ideal voice restoration therapy that combines low invasiveness, higher speech intelligibility, affordability, and accessibility is essential. Machine learning algorithms, particularly deep neural networks (DNNs), have the potential to achieve these four parameters by predicting speech from a variety of different markers and converting the predicted phonemes to intelligible, computer-generated speech. Neural networks in a nutshell are composed of an input layer, output layer, and a variable number of hidden layers; DNNs are characterized by the presence of at least two hidden layers. The network accepts an input from a training set where the correct output is already known and predicts the output by using calculations in each layer. The predicted output is compared to the known output, and the error rate is used to adjust the calculations that occur in the hidden layers. This occurs for a preset number of iterations determined by the user with the goal of getting the predicted output as close as possible

to the known output. Once the DNN is trained, it can be used to predict values for data that does not have a known output. This project seeks to create and train a multi-pipeline DNN consisting of several convolutional neural networks (CNNs) and long short-term memory (LSTM) networks. The DNN will take both visual and auditory signals from videos of laryngectomees and non-laryngectomees, with and without an electrolarynx, to create a computer-generated, intelligible, “normal” voice with a wide range of volumes. Since this algorithm builds on the existing electrolarynx, it retains its non-invasive qualities. However, the wide range of volumes will greatly increase speech intelligibility, even in noisy settings. The creation of this algorithm comes with the intention to eventually implement the algorithm in a smartphone application, which will increase both the affordability and accessibility of the voice restoration therapies. This DNN has the potential to increase the quality of life of laryngectomees by increasing the extent of which they can communicate.

Prior Research and Innovation

While this project seeks to combine both visual and audio data to perform speech analysis, current research focuses primarily on either one or the other modality. Current research in artificial lip reading using visual data focuses on the implementation of artificial neural nets (ANNs) to achieve higher accuracy than human lip reading. LipNet used a combination of a convolutional neural network (CNN) and a recurrent neural network (RNN) to learn phonemes based on both spatial and temporal features. It improved on previous designs by combining feature extraction and prediction into a single pipeline rather than two wholly separate networks. It also predicted full sentences at a time rather than words. LipNet had a higher accuracy score than human lip-readers, which points to the implementation of deep learning models as the future of lip reading¹⁴. Another study used a Microsoft Kinect camera to track seventeen different points on the lip. This input was fed into an ANN and used to predict short phrases. The accuracy of this method was 77.2%, which is lower than other models. However, this study emphasized using low cost, portable materials, which is a step in the right direction for improving voice restoration methods¹.

Efforts to improve the electrolarynx have come from multiple different directions, mostly including the attenuation of generic vibrational noise from the electrolarynx, which detracts from the intelligibility of the actual speech, and creating a pitch range that more adequately represents the pitch of the voice of patients. Padmini proposed the use of Mel frequency cepstral

coefficients (MFCC) to extract essential features from electrolarynx output. The frequency of the electrolarynx would not be extracted with these essential features, thus attenuating generic vibrational noise⁸. Syrinx is a wearable artificial larynx that expands the frequency range of a traditional electrolarynx. Besides having an increased frequency range for the electrolarynx, Syrinx uses recordings of the user’s voice to determine a personalized vibrational frequency. Finally, the wearability of this artificial larynx solves the problem of needing an extra hand to operate the electrolarynx⁹. A different approach to the electrolarynx is the use of electromyography, where magnets are embedded in the mouth to capture electromagnetic data that predicts speech. This is obviously both invasive relative to the electrolarynx and uncomfortable.

This project expands on the research above by creating an DNN that utilizes both video processing and audio processing rather than a single modality to predict speech as accurately as possible. Since lip movements and audio output from the electrolarynx are both very different data modalities, the project will be split into two separate pipelines. Each pipeline will predict a phoneme for that respective modality separately; the two pipelines will then be combined using ensemble learning for the final speech prediction.

Materials and Methods

Data Collection

The given videos were supplied by Rachel Jonas, M.D. of the ENT department. They were pre-recorded by advisors to avoid extra IRB approval. Five participants of both sexes were included and per the IRB, all other information barring sex were excluded. To further preserve the anonymity of the participants, the upper half of the face was cropped out of the frame for all videos. The number of participants was chosen to be five due to the use of five participants in the Xbox Kinetic Speech program as described in the Introduction¹. The participants chosen are all English-speaking adults with functional larynges without diagnosed voice or speech pathology recruited personally by the designers of the study. Each speaker repeated the first three sentences of the *Rainbow Passage* while audiovisual output was recorded. After training by Holly Hess, a certified speech-pathologist, the participants re-read the passage using an electrolarynx while articulatory speech movement and acoustic output were again captured. The *Rainbow Passage* was chosen because it is short and phonetically balanced. In total, there are 52 electrolarynx based files and

51 normal speaking files. The same participants were then used to record videos of conversational speech using the same methodology. Each participant had freedom to choose what they wanted to say for sixty seconds. All of these videos are located on a shared UVA dropbox folder. During this process, a Rivanna allocation, *uvavoice*, was purchased due to the high volume of computation necessary to train a DNN on high-dimensionality data such as images and audio clips.

Audio Pipeline

All the audio processing was done using Python 3.9. The audio was extracted from the videos using the *scipy* package on Python, which contains a module for signal processing. This *wavfile* module converted the videos into audio files with the *.wav* extension. The electrolarynx audio files were processed using a Butterworth filter, defined using modules in the *scipy* module, in order to remove background noise. The argument for the butter bandpass filter requires arguments for a low frequency, high frequency, and a cutoff frequency. The lower and higher limits were determined using trial and error using one recorded video per participant. These limits appeared to occur at 500 and 2000 Hz, respectively.

After noise removal, Mel Frequency Cepstral Coefficients (MFCCs) were initially created for all the videos also using the *surfboard* package, an open-source library for audio processing¹⁵. MFCCs were chosen to act as features for the

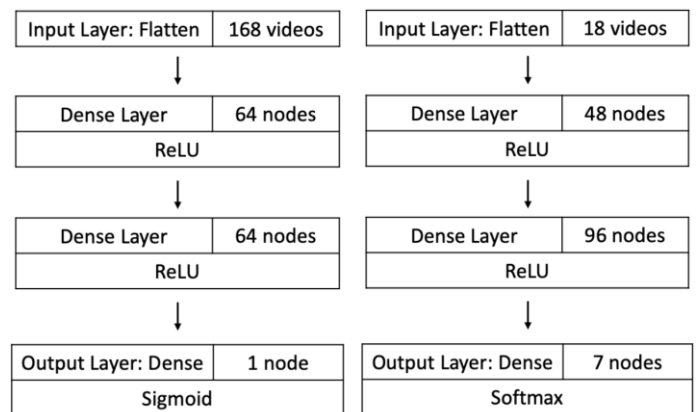


Fig. 1. DNN for Binary Classification (left) and Multi-label Classification (right). The images above describe the layers in each DNN including the number of nodes and the layer’s respective activation functions. For the binary classification DNN (left), there are 4 total layers, with the hidden layers each having 64 nodes and a ReLU activation function. The output layer, since its binary classification, only has 1 node and a sigmoid activation function. Similarly, the multi-label classification DNN (right) has a total of 4 layers with 2 hidden layers each with ReLU as the activation function; however, the number of nodes doubles as you go from layer 2 to layer 3, 48 and 96 nodes, respectively. In its output layer, there are 7 nodes since there are 7 possible labels, and since it is multi-label classification, a softmax activation function is used.

audio data because they are a popular spectral-based parameter used in audio processing. They represent the frequencies surrounding the human vocal range, where the values of the MFCCs amplify this range to improve clarity¹⁶. The matrices of coefficients, known as cepstral coefficients, contain information about the rate changes in the different spectrum bands. Positive coefficients indicate that the spectral energy of the audio is concentrated in low-frequency regions. If the cepstral coefficients have a negative value, the spectral energy is concentrated in high-frequency regions¹⁷.

Binary Classification Using MFCCs

In order to determine whether the MFCCs created useful features to distinguish between electrolarynx and non-electrolarynx audio clips, binary classification was performed using a deep neural network (DNN). MFCCs for this task were generated using *librosa*, a python package designed for music and audio analysis. The neural network was created using the *keras* library and a Sequential model. The DNN consisted of an input layer, two hidden layers, and an output layer (Figure 1). The input consisted of 168 audio files, which were first passed into a flatten layer. The flatten layer prepares data to enter the neural network. The two hidden layers in this network consisted of dense layers with 64 nodes each. This number of nodes was chosen because 2ⁿ nodes show the best results in neural networks. These dense layers contain full connectivity between the previous layer and the current one. ReLU was chosen as the activation function in these layers because it is known to overcome the vanishing gradient problem and allow the model to learn faster and perform more efficiently. Finally, the neural network contained an output dense layer containing one node with a sigmoid activation function. One node was used because the goal of the network is binary classification. Since the sigmoid function exists only between 0 and 1, it performs well during binary classification (Figure 1).

Multi-Label Classification Using Singular Words

Since previous literature tends to classify animal sounds, single words, or emotions, the audio files were split into singular words from the beginning of the first sentence of *The Rainbow Passage*. The words include *when, the, sunlight, strikes, raindrops, in, air*. Audacity was used to manually split the audio files into separate words. These words were then passed into another DNN consisting of an input layer, two dense layers, and an output layer. The input flatten layer processed 18 audio clips into the network. The two hidden layers consisted of dense layers with 48 and 96 nodes, respectively, and ReLU as their activation function.

The output dense layer consisted of 7 nodes with softmax as the activation function because there were 7 possible words. Softmax was chosen as the activation function because it is commonly used in multi-label classification. The network predicted which word the audio file contained (Figure 1).

Video Pipeline

The entirety of the video pipeline consisted of using a software called DeepLabCut to perform initial feature extraction from each of the *Rainbow Passage* videos followed by a convolutional neural network (CNN) that reduced the dimensionality of the image to four points and a long short-term memory network (LSTM) that predicted phonemes from lip positions. The output of the LSTM would be weighed in the final phoneme prediction after combination of the audio and video pipelines.

Feature extraction using DeepLabCut

DeepLabCut is a pose estimation software developed by the Mathis Lab that has been traditionally used to model small mammal movement but has the potential to be used to predict movement of body parts in humans as well. The software uses a representative set of labeled frames as training data to predict the specified labels in unlabeled videos. Thus, a small number of representative frames can be used to determine labels in long videos with thousands of frames. This results in the reduction of the dimensionality of each frame from thousands of pixels to the small number of labels, which can greatly reduce the computational expense of each frame in future neural networks (Figure 2).

For this project, a total of four points on the lip were labeled: the upper, lower, left, and right corners of the lip (Figure 2). This number was chosen based on prior research with the intention to minimize computational expense while

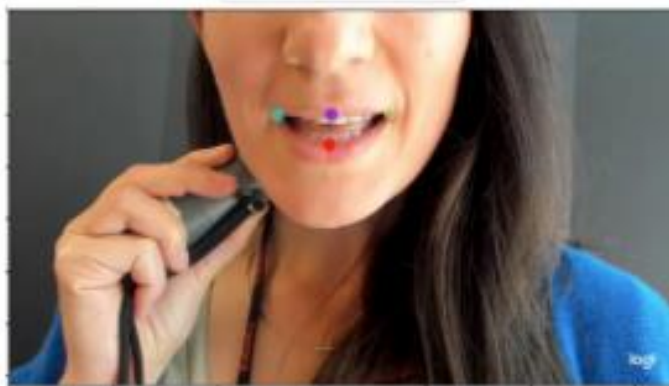


Fig. 2. DeepLabCut Labeled Frame. To get the labeled frame as seen above, DeepLabCut trains a ResNet CNN on manually labeled frames and then uses the trained model to make predictions on the upper, lower, left, and right corners of the lip.

maximizing the potential for accurately predicting the test set. Using the software, each *Rainbow Passage* video was split into several thousand frames, with the exact number of frames depending on the video. The frames were left uncropped since the speakers assumed a distance from the camera that was indicative of our final application. For each video, these frames were clustered into 50 clusters using k-means clustering; the number of clusters was chosen to represent the 44 phonemes in the English language as well as several extra clusters to represent the possibility of slightly different mouth positions for any given phoneme. Due to the considerable time needed to manually label 50 frames per video, five videos with an electrolarynx and five videos without an electrolarynx for each participant were randomly selected to be labeled using a random number generator. These 2500 frames were manually labeled with the four points on the lip. The frames served as the input for the next step in the pipeline, the CNN.

ResNet CNN

DeepLabCut has multiple options for which CNN is used to predict the lip positions; ResNet-50 was initially chosen due to the lower number of layers and faster training time. ResNet-50 has three defining features: convolutional layers, pooling, and fully connected layers. The 50 convolutional layers are used to extract the most important features of the image; the dot product of a small matrix called the kernel and small pixel sections of each image is found for all parts of the image in order to perform this step. Pooling is characterized by finding either the average or maximum value of a set of pixels in the image; this is done to reduce dimensionality of the image. A final fully connected layer with 1000 nodes resembles a normal neural network that classifies that reduced image. This classification was used to determine where each of the four points on the lip were for that particular image.

This ResNet-50 model was trained on the 2500 labeled frames. To test for possible overfitting, or low error rates for training data but high rates for testing data, the remaining unlabeled frames from the *Rainbow Passage* videos that were not selected were used as a validation set. To ensure the usefulness of DeepLabCut before moving on to the next step in the pipeline, the model was tested on the recorded videos of conversational speech. Finally, to determine the necessity of a uniform speaker depth from the camera, the model was tested on frames from another open-source dataset, where the speakers were positioned farther away from the camera.

DeepLabCut utilizes an error measurement called mean pixel error to determine the accuracy of the model. It does this by finding the mean distance in pixels between the predicted lip position labels and the actual lip position labels. A mean pixel error of less than 5 pixels is recognized to be an appropriate threshold for determining if the model can accurately predict the labels; this threshold was chosen as it is an approximate for pixel error in low-resolution cameras. The mean pixel error was recorded for each round of videos; final pictures and videos of predicted lip positions were also qualitatively judged for accuracy.

Long Short-Term Memory Network

The final step in the video pipeline was a long short-term memory network (LSTM) that took a set of lip positions for each video and predicted a phoneme for each frame. LSTM networks are unique in that they retain useful information about past data and use this information to inform the prediction of the current data. This quality makes LSTM networks particularly useful for speech recognition, where there are many strings of consecutive phonemes that occur frequently in English speech. The LSTM model would be implemented using PyTorch with an input size of 4 (representing each of the four labeled points on the lip) and an output size of 1 (representing the predicted phoneme).

The size of the hidden layer would be adjusted based on the accuracy of the model.

Similar to the testing of the accuracy of DeepLabCut, the LSTM model would then be trained on the lip positions from the *Rainbow Passage* videos and tested on the conversational speech videos. Further testing includes using videos of laryngectomees speaking the *Rainbow*

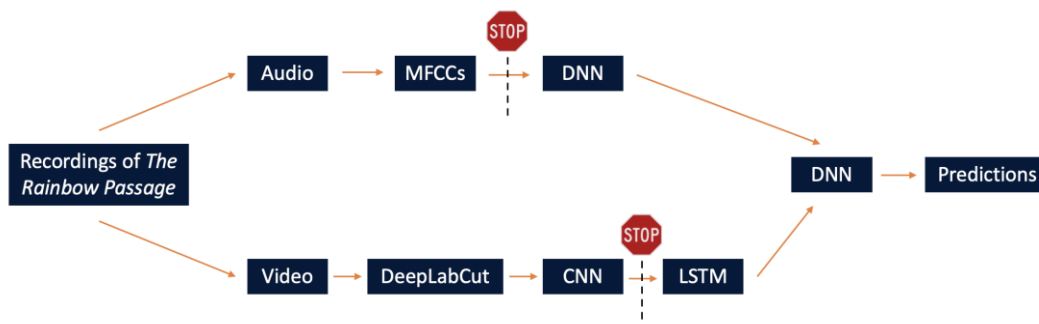


Fig. 3. Overall pipeline of project. This image shows the overall pipeline including the separation of the audio and video portions. The stop signs indicate what was achieved in the scope of this time.

Passage as well as speaking conversationally to ensure that the model is still effective for laryngectomees. For each video, the accuracy score of the model at correctly classifying phonemes will be determined for a variety of epoch lengths and hidden layer sizes.

Pipeline Combination

The audio and video pipeline each output a predicted phoneme for each frame of each video, which must then be analyzed and combined to predict a final phoneme. Since the audio and video pipelines could theoretically output different phonemes for a given frame, some form of ensemble learning is required to accurately predict the most correct phoneme for any given frame. Since the audio and video pipelines are separate throughout the entire process of predicting a phoneme from a given modality, the two can be viewed as separate prediction models, with one having the potential to be more accurate than the other.

The solution to combining the two pipelines was to use a boosting algorithm. Boosting involves iteration through each classifier, determining the accuracy of each classifier, and weighting that prediction based on the accuracy of that classifier. AdaBoost will be used to perform this step as it is a common and familiar boosting algorithm that can be easily implemented in Python using scikit-learn. The accuracy of the final predicted phoneme will be measured and compared with the accuracy of the individual classifiers. If the accuracy is relatively low, the possibility of increasing the number of classifiers by iterating through the same classifier multiple times will be explored to improve the accuracy of our novel ensemble program (Figure 3).

Results

Audio Pipeline

Binary Audio Classification

The primary goal for audio classification was to be able to distinguish the *Rainbow Passage* sentences repeated by the participants. After realizing that training with the whole video and extracted MFCC's was not feasible, it was determined that it would be important to see whether the MFCC's provided valuable information in distinguishing between larynx and electrolarynx speech. Using this approach and the idea of binary classification deep learning, *Rainbow Passage* videos were labeled as either normal or electrolarynx, and the features for each was the matrix of coefficients given by the MFCCs. Using the 4-layer DNN described above, this binary classification yielded a training

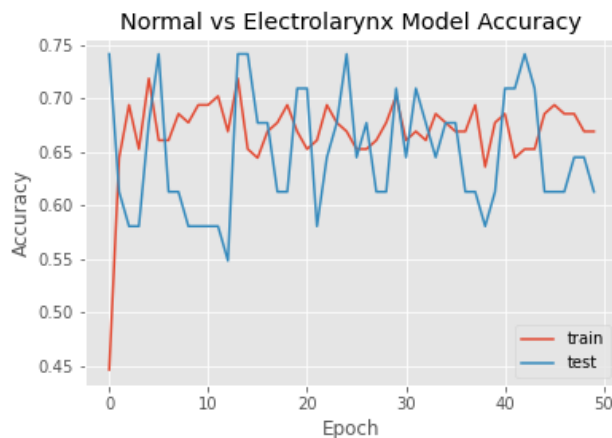


Fig. 4. Accuracy Graph for Binary Classification. This graph shows the test and train accuracy for the binary classification DNN over 50 epochs. The jump-like tendencies of the graph are also another indication of random prediction and not model learning.

accuracy of 69% and a testing accuracy of 61%. Looking at the cross-entropy values, the training cross-entropy was 0.59 and the testing cross-entropy was 0.73. Cross-entropy is one example of a loss function that describes how well the algorithm, whose output is a probability value between 0 and 1, is classifying. Also known as log loss, the cross-entropy loss increases as the difference between the predicted and actual label becomes larger. Ideally, a log loss of 0 correlates to a perfectly accurate model¹⁸. In this case, with both values being so small, it indicates that the classification being done by the model is essentially as random as a coin-toss, meaning that no learning is occurring (Figure 4).

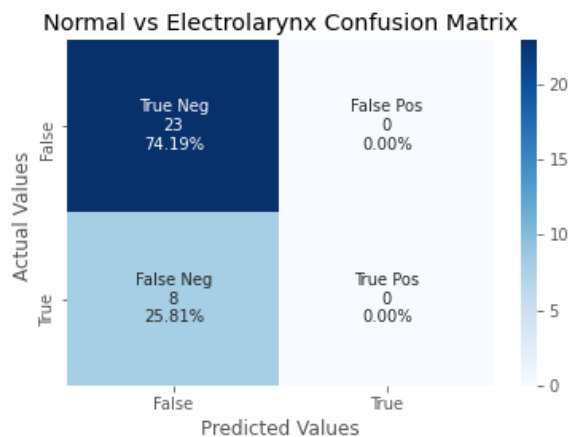


Fig. 5. Confusion Matrix for Binary Classification. This graph shows the confusion matrix represented by the algorithm's predictions. Ideally, since actual predictions are plotted on the vertical axis and predicted values are plotted on the horizontal axis, the diagonal should be the darkest shaded since those represent areas where the actual and predictive values match. However, this confusion matrix has the highest values in the true negative and false negative sections rather than the true negative and true positive sections. This highlights how the network is not able to distinguish between normal and electrolarynx.

To further assess the classification of the DNN designed, a confusion matrix was generated. The confusion matrix, which plots actual labels against predicted labels, classified the test dataset into one of the four categories as shown in the figure below. For this particular classification, there were 23 true negatives and 8 false negatives, reaffirming that the algorithm was not actually learning, and that the prediction was happening somewhat randomly. Ideally, the highest values should be seen in the true negative and the true positive labels because these indicate correct predictions (Figure 5).

Multi-Label Audio Classification

As another thought experiment, a new DNN was developed with the intent of performing multi-label classification with the labels: when, the, sunlight, strikes, raindrops, in, air. This is the first part of the first sentence in the rainbow passage. The idea stems from literature published on audio classification with noisy data. Due to the additional noise, the most basic approach to take is to determine whether an algorithm can differentiate between singular words. Using this literature and the DNN based on the idea of multi-label classification, the training accuracy was 17% and the testing accuracy was 15%. The training cross-entropy was 2.33 and the testing cross-entropy was 2.48. Because the data only consisted of 18 videos, no immediate conclusions can be drawn from these results (Figure 6).

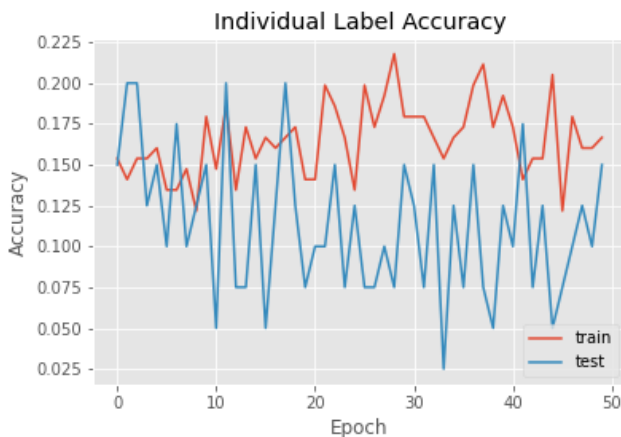


Fig. 6. Accuracy Graph for Multi-label Classification. Similar to the accuracy graph of the binary classification model, this graph shows the test and train accuracy over 50 epochs for multi-label classification. The repetition of the spike-like pattern of the graph indicates that the algorithm is not truly learning and that its predictions are quite random. Due to the small size of the dataset, no real conclusions can be drawn.

Video Pipeline

Rainbow Passage and Open-Source Data Set

The first aim of this project was to design an artificial neural network by mapping English phonemes to visual and

acoustic electrolarynx in individuals with functioning larynges using a prescribed reading passage. The first step in this process was to create a visual speech recognition network to track the movements of a subject's lips as they were speaking the rainbow passage. The rainbow passage

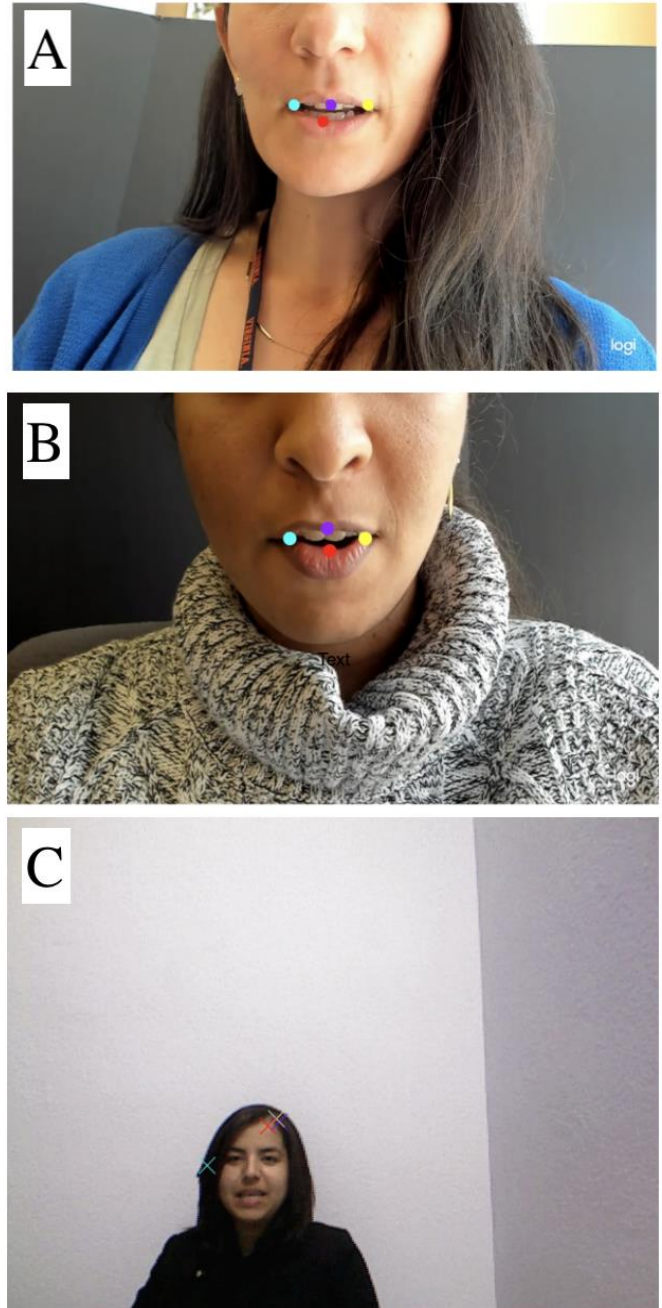


Fig. 7. Rainbow Passage and Open-Source Data Set Test and Training Frames. (A) A labeled training frame from the rainbow passage videos. The ResNet CNN is labeling the four points on the lip. (B) A labeled test frame from the rainbow passage video where the ResNet CNN is predicting and labeling the 4 points on the lip. (C) A labeled test frame from the open-source data set where the trained model is not accurately predicting and labeling the points on the lip but on the subject's forehead and hair.

videos were used to train the ResNet CNN created by DeepLabCut because it contains all the English phonemes. DeepLabCut automatically splits the 2500 labeled rainbow passage video frames into a training and test set. 95% of the frames were used for training and the remaining 5% were utilized in the test set. After DeepLabCut created the training set, DeepLabCut trained on the labeled frames using the ResNet CNN with 500,000 iterations. After training on the rainbow passage was completed, DeepLabCut evaluated the trained ResNet CNN on the test set. As shown in Table 1, the mean pixel error for the training, 4.19 pixels, and test are close to five pixels meaning the trained model is accurately predicting the 4 points on the lips. Additionally, as seen in Figure 7, the model is accurately predicted and labeled the 4 points on the lips of the participants. The test mean pixel error was greater than training pixel error due to the number of frames relegated to training compared to testing.

Table 1. Mean Test Pixel Errors for Test Data. The above table displays results of the mean test pixel error for all three test data sets. The open-source data set had the highest mean pixel error. Both rainbow passage and conversational speech had mean pixel errors close to 5 pixels. The trained ResNet CNN was used to test on the data sets.

Data	Mean Test Error (pixels)
Rainbow Passage	5.22
Open-Source	134.65
Conversational Speech	5.11

After the trained ResNet CNN was validated using the test rainbow passage frames, the trained network was used to predict the 4 points on a lip on an open-source data set provided by our advisors. The open-source data set was used as another validation test set in order to see whether the trained network could predict points on videos that were not specifically recorded for this project. As seen in Table 1, the mean test pixel error for the open-source data set is



Fig. 8. Conversational Speech Frame. The above figure represents a frame from the conversational speech test data set. The four points on the lip represent the trained ResNet CNN's labeled prediction. The model is making accurate predictions for all 4 points.

much greater than 5 pixels meaning the model is not accurately predicting the points on the lips. Furthermore, as seen in Figure 7, the model was predicting the 4 points close to the forehead of the participant rather than the lips. The lack of accuracy is attributed because the lips of the participants in the training videos were at a closer distance to the camera compared to the open-source data set.

Conversational Speech

The second aim of this project was to apply a trained ANN to conversational speech in laryngectomees. However, to achieve this aim, the trained ResNet CNN was first applied to videos of conversational speech recorded by participants in order to see whether the model can predict points on the lips on conversational speech. Conversational speech videos were used because conversational speech does not utilize all the phonemes and syllables in the English language. Again, as shown in Table 1, the mean pixel error for the conversational videos is close to five pixels meaning the trained model was accurately predicting the four points on the lips. Moreover, as depicted in Figure 8, the ResNet CNN is accurately labeling the four points on the lips of the participant during conversational speech.

Discussion

The goal of this project was to create a DNN combining audio and visual analysis of individuals with and without the electrolarynx. The DNN was meant to predict phonemes in the English language. Since MFCCs are a very common method to extract audio features, it was the method used for the audio pipeline. DeepLabCut was used to identify four points on the lip for the video pipeline. Although the results from the audio portion were lacking, the video portion showed a high level of accuracy.

In order to improve the audio portion, new methods such as a Melspectrogram, Spectral Contrast, and Wavelet transforms can be used as features. Recently, more research has been performed using these methods for training neural networks instead of MFCCs. Artificial Intelligence has also shown more promise learning from images than MFCCs in this area of research. The methods listed above would produce images containing the information from the audio files similar to how MFCCs data can be displayed on a graph. Additionally, the multi-label classification performed with singular words showed a lot of promise. Since the clips containing singular words are shorter, it is easier for the neural network to learn this data. With singular words, the possibility for the words is also finite, allowing supervised learning to be more accessible. With a

small bank of representative words in the English language, models could be trained with singular words rather than phrases, sentences, or paragraphs, which also take more processing power. After learning individual words, the model could graduate to more complex structures similar to how language processing has evolved.

The next step in the video pipeline would be to associate groups of frames with labeled points to different phonemes or syllables. This process would also be facilitated with videos of singular words rather than several sentences, especially if the words are one syllable. Once the neural network has trained on simple syllables and achieved high accuracy, harder words can be added. More research needs to be performed regarding how to differentiate between visual sounds in the English language. For example, b and p, although audibly different, look very similar on video because of the positioning of the lips.

Since the goal of this project is to create a device or an app that electrolarynx users can operate to improve their speech intelligibility, the diversity of the training data needs to increase. Currently, the training data includes only five subjects who speak English, resulting in only about 200 videos. Machine learning models, in general, require thousands of inputs in the training set to reach high levels of accuracy and applicability. With only five subjects, the variety of ethnicities and accents represented by actual electrolarynx users is not present in the data. Since the videos only contain English words, this project is also not applicable to other languages and cannot be used fully by multilingual electrolarynx users or international users. Furthermore, all the videos are taken at the same distance from the face. Realistically, electrolarynx users might hold their phones at varying distances when using the app or device. Therefore, more data needs to be collected and trained to encompass a larger range of distances from the face and overcome other weaknesses in the current dataset.

End Matter

Author Contributions and Notes

Sameer Agrawal and Katherine Taylor worked on the video pipeline and contributed to writing the paper. Surabhi Ghatti and Medhini Rachamalla worked on the audio pipeline and contributed to writing the paper. The authors declare no conflict of interest.

Acknowledgments

Timothy Allen, PhD
Shannon Barker, PhD

Noah Perry
Vignesh Valaboju
James Daniero, M.D.
Haibo Dong, PhD
Rachel Jonas, M.D.
John Kelly
Jessica Lin
Holly Hess, SLP

References

1. Kohlberg, G. D., Gal, Y. (Kobi) & Lalwani, A. K. Development of a Low-Cost, Noninvasive, Portable Visual Speech Recognition Program. *Ann Otol Rhinol Laryngol* **125**, 752–757 (2016).
2. Antin, F., Breheret, R., Goineau, A., Capitain, O. & Laccourreye, L. Rehabilitation following total laryngectomy: Oncologic, functional, socio-occupational and psychological aspects. *European Annals of Otorhinolaryngology, Head and Neck Diseases* **138**, 19–22 (2021).
3. Wulff, N. B., Højager, A., Wessel, I., Dalton, S. O. & Homøe, P. Health-Related Quality of Life Following Total Laryngectomy: A Systematic Review. *The Laryngoscope* **131**, 820–831 (2021).
4. Andrews, J. C., Mickel, R. A., Monahan, G. P., Hanson, D. G. & Ward, P. H. Major complications following tracheoesophageal puncture for voice rehabilitation. *The Laryngoscope* **97**, 562–567 (1987).
5. Albirmawy, O. A., Elsheikh, M. N., Saafan, M. E. & Elsheikh, E. Managing problems with tracheoesophageal puncture for alaryngeal voice rehabilitation. *The Journal of Laryngology & Otology* **120**, 470–477 (2006).
6. Wang, R. C. *et al.* Long-term Problems in Patients With Tracheoesophageal Puncture. *Arch Otolaryngol Head Neck Surg* **117**, 1273–1276 (1991).
7. Carr, M. M., Schmidbauer, J. A., Majaess, L. & Smith, R. L. Communication after laryngectomy: An assessment of quality of life. *Otolaryngol Head Neck Surg* **122**, 39–43 (2000).
8. Padmini, L. Neural Network based New Bionic Electro Larynx Speech System. *International Journal of Engineering Research* **5**, 5 (2017).
9. Takeuchi, Masaki. Information about Syrinx. (2021).
10. Rameau, A. Pilot study for a novel and personalized voice restoration device for patients with laryngectomy. *Head & Neck* **42**, 839–845 (2020).

11. Gates, G. A. *et al.* Current status of laryngectomy rehabilitation: I. Results of therapy. *American Journal of Otolaryngology* **3**, 1–7 (1982).
12. Souza, F. G. R. *et al.* Quality of life after total laryngectomy: impact of different vocal rehabilitation methods in a middle income country. *Health and Quality of Life Outcomes* **18**, 92 (2020).
13. Cox, S. R. & Doyle, P. C. The Influence of Electrolarynx Use on Postlaryngectomy Voice-Related Quality of Life. *Otolaryngol Head Neck Surg* **150**, 1005–1009 (2014).
14. Assael, Y. M., Shillingford, B., Whiteson, S. & de Freitas, N. LipNet: End-to-End Sentence-level Lipreading. *arXiv:1611.01599 [cs]* (2016).
15. Lenain, R., Weston, J., Shivkumar, A. & Fristed, E. Surfboard: Audio Feature Extraction for Modern Machine Learning. in *Interspeech 2020* 2917–2921 (ISCA, 2020). doi:10.21437/Interspeech.2020-2879.
16. Suksri, S. *Speech Recognition using MFCC*. (2015). doi:10.13140/RG.2.1.2598.3208.
17. Hossan, Md. A., Memon, S. & Gregory, M. A. A novel approach for MFCC feature extraction. in *2010 4th International Conference on Signal Processing and Communication Systems* 1–5 (2010). doi:10.1109/ICSPCS.2010.5709752.
18. Ho, Y. & Wookey, S. The Real-World-Weight Cross-Entropy Loss Function: Modeling the Costs of Mislabeling. *IEEE Access* **8**, 4806–4813 (2020).

Supplemental Information

Link to Github

<https://github.com/ghattisurabhi/Electrolarynx-Capstone>