UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

DEVELOPING AN ALGORITHM INTEGRATING VOICE AND IMAGING ANALYSIS TO RECOGNIZE FACIAL FEATURES AND DEFICIENCIES AFTER ORAL SURGERY

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

MASTER OF SCIENCE

in

Electrical and Computer Engineering

By

ERFAN SEIFI
Norman, Oklahoma
2024

DEVELOPING AN ALGORITHM INTEGRATING VOICE AND IMAGING ANALYSIS TO RECOGNIZE FACIAL FEATURES AND DEFICIENCIES AFTER ORAL SURGERY

A THESIS APPROVED FOR THE

SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

BY

Dr. Hazem Refai, Chair

Dr. Joseph Havlicek

Dr. Choon Yik Tang

© Copyright by ERFAN SEIFI 2024
All Rights Reserved.

Acknowledgments

I would like to express my deepest gratitude to my advisor, Dr. Hazem Refai, for his invaluable guidance, patience, and support throughout the time spent pursuing this degree. His expertise and insightful feedback have been instrumental in shaping both this research and my development as a scholar. I would also like to extend my sincere thanks to the members of my committee, Dr. Joseph Havlicek and Dr. Choon Yik Tang, for reviewing this thesis. Furthermore, I would like to thank all my professors, peers, mentors, friends, and staff including, Ms. Denise Davis at the University of Oklahoma, for their valuable support throughout my studies.

Finally, I would like to thank my parents and sister for all their support and patience during my graduate education.

Table of Contents

List of Tables

List of Figures

Abstract

According to the National Institute of Health (NIH), oral cancer is one of several major types of head and neck cancer (HNCs) and affects approximately 54,000 individuals in the United States each year [20]. Recognized risk factors for HNCs are primarily tobacco use, alcohol intake, and inadequate oral hygiene, the latter of which is significant for oral cavity cancer [22, 23, 24]. Like treatment for other cancers, oral cancer therapies usually include surgery, radiotherapy, chemotherapy or a combination thereof [21, 25, 26, 27]. Treatments can cause loss of clear speech as a result of resecting parts of the vocal tract, which alters the vocal tract shaping and/or limits mouth movement.

For this thesis, a software application was developed to evaluate a participant's spoken communication by simultaneously analyzing facial features and voice recordings of him or her reading a scripted passage. The effect of vocal tract changes following oral surgery was investigated using the new application, which showed measurable, quantifiable loss of speech. The goal of development and testing was providing medical doctors, speech therapists, and researchers the ability to leverage data-drive algorithms when designing strategic rehabilitation treatment plans to improve patient recovery. With the use of machine learning techniques, a model was developed for analyzing speech patterns and identifying/quantifying an emulated impact of oral surgery on generating speech. Such an approach leverages acoustic analysis and offers a non-invasive, accessible means of assessment, especially when compared to other methods (e.g., high-speed video-stroboscopy) that are known to cause side effects of swelling/pain and exclude some cancer patients. By focusing on extracting and analyzing various audio features from speech recordings and spatial dynamics of the lips—including formant frequencies—investigators are able to discern

subtle changes in motor speech task characteristics. This information could indicate post-surgical complications or suggest improvements during recovery. The framework built in this thesis identifies a process for comparing speech samples and special facial dynamics both before and after surgery. Detecting impairments, like shift in speech frequencies, offers valuable feedback about a patient's motor speech task monitoring and rehabilitation progress. Results demonstrate the effectiveness of therapeutic interventions after cancer treatment.

Experimental analyses emulated possible post-surgical scenarios for two healthy participants. The first participant was a non-native English speaker and the second participant was a native English speaker with American accent. Various speech patterns were observed under both regular conditions and those experienced as a consequence of two types of oral obstructions. Preliminary results demonstrate the potential for using the novel method detailed herein for objectively assessing speech loss and monitoring speech rehabilitation for patients who suffer from oral cancer. In short, this thesis presents a framework for non-invasive assessment of speech impairments following oral cancer treatment that bridges the gap between clinical speech therapy and computational speech analysis. The impact will enhance oral health and surgery rehabilitation.

This study was conducted under an approved IRB by the University of Oklahoma No. 17042 and title: AI For Facial Rehab Post Oral Surgery Speech Recovery.

1 Introduction

The ability to accurately analyze and interpret human speech has profound implications across a wide range of fields, from healthcare and assistive technologies to communication and artificial intelligence. Speech motor task is one of the most natural and fundamental means of human expression and communication, carrying not only linguistic information but also nuances that convey emotions, intentions, and even the speaker's physical condition. In healthcare, for instance, speech analysis can offer non-invasive diagnostics and monitoring for conditions affecting speech capabilities, such as those experience in the aftermath of oral surgery. By focusing on the differentiation between regular and emulated speech—along with cross-subject analysis, the study undertaken for this thesis aims to contribute to the broader understanding of how speech characteristics vary under different physical restrictive conditions of the oral cavity, as emulated through post oral-operative data for a variety of individuals. The methodology is suggested for clinical diagnostics, post-operative care, and rehabilitative therapy to track speaking rehabilitee and improvement during speech therapy.

The proposed method integrates voice and imaging analysis to recognize facial features and speech deficiencies after oral cancer surgery. Results provide a way to characterize the loss of speech motor task ability by utilizing recorded audio and visual data gathered while a participant reads a scripted passage both before and after surgery. Data was collected using a commercial camera to capture the act of two participants reading a passage from the novel “the caterpillar” [30]. Motor speech disorders simulated by two emulated conditions, namely oral obstacle under the tongue and clenching teeth together while speaking were assessed to a) mimic oral surgery effects and b) limit oral cavity motion. Findings were compared with baseline speech data gathered under normal conditions (i.e., not restrictions or oral obstructions during speaking). Prior to processing, the audio

and video signals were separated. Images extracted from the video were processed using artificial intelligence (AI) to identify facial features (i.e., landmarks) on the participant's face, including lip, nose, eyes, and face boundary. An algorithm was then implemented to track and synchronize lip movements with the audio signal. Next, passage words and their associated timings were separated using a speech recognition algorithm so that corrections to the participant's lateral head movements that might negatively impact lip tracking accuracy could be made. Audio-visual features, including formants, pitch, vertical and lateral lip opening/closing displacements and rate, among other parameters, were calculated for each word uttered by the participant while reading the passage. Features were subsequently utilized to determine post treatment speech loss and to quantify recovery and gain-to-normal changes during speech therapy.

Method validation was achieved using data gathered from two healthy participants reading a passage "the caterpillar."—Each participant was evaluated during normal conditions and also under two restrictive, emulated conditions: 1) speaking with a hard candy in the mouth and 2) speaking with clenched teeth. Preliminary results demonstrate the potential of the proposed method for objectively assessing speech loss and monitoring speech rehabilitation for oral cancer patients. Machine learning models were developed to differentiate normal speech with restrictive speech for an individual with an 80% F1 score. The results underscore the efficacy of integrating voice and imaging analysis when assessing and monitoring speech rehabilitation after oral cancer surgery. The method's ability to objectively quantify speech loss and to track the speech task rehabilitation progress via audio-visual feature analysis represents a significant advancement in personalized cancer care. Compared to traditional subjective assessments of speech quality, the novel approach detailed in this thesis offers a more precise, data-driven evaluation and specifically targeted interventions.

1.1 Thesis Contributions

This thesis analyzes simultaneous lip motion patterns and speech audio signals that both characterize the loss and track the improvement of speech therapy for patients undergoing oral cancer treatment. Method includes image alignment and processing, digital signal processing, and statistical analysis, as well as the use of machine learning classification models. Findings outlined in this study enhance patient recovery experiences after oral surgery and aid both speech therapists and researchers when selecting/applying more strategic techniques and approaches for minimizing recovery time and reducing medical costs.

Primary contributions of this study are summarized below.

[1] Implemented an automated process to measure the displacements of the facial expressions including lip movement while correcting for head motions.

[2] Implemented an automated framework to measure audio features including zero crossing, formant frequencies, and pitch frequency per word.

[3] Constructed a machine learning algorithm to differentiate between a baseline audio-visual recording and a recording of passage reading inhibited with a hard candy in the mouth.

[4] Evaluated the importance of each machine learning feature to detect drift from the audiovisual baseline.

2 Related Work and Background

Head and neck cancer (HNC) impacts over half a million people worldwide; its management exerts a significant burden on the patient and the healthcare system [16]. Impairments after final treatment of HNC could include physical appearance, speaking ability, swallowing, chewing, saliva production, nerve-movement connectedness, discomfort, and dietary health [15]. The authors in [4] focused on evaluating the impact of HNC treatments on patients' functional outcomes and quality of life, using quality of life questionnaires. Specific activities assessments included eating, swallowing, speaking, social participation, and pain management. Results indicate treatments like radiation therapy, surgery, and chemoradiotherapy can lead to significant functional impairments and reduced quality of life. Authors advised that understanding patients' priorities and perspectives is crucial for treatment plan evaluation and decision-making.

Speech and swallow rehabilitation for HNCs includes several steps and guidelines. Clinical guidelines outlined in [1] suggest the importance of completing an evaluation of patients' speech and swallowing abilities before treatment to establish a baseline. The guideline emphasizes that patients undergoing radiation for HNC often experience trouble swallowing due to side effects, like soreness and swelling.

Findings by authors in [3] stress the need for a focus on trismus (i.e., limited mouth opening) for postoperative management, especially for patients with oral and oropharyngeal cancers, to improve their rehabilitation and quality of life. Limited mouth opening is a common and significant complication that hinders activities like eating, drinking, and speaking, thus patient quality of life. Their study examined 101 patients who filled out a questionnaire focusing on nutritional, sensual, and speech disorders, along with pain. Researchers evaluated maximal interincisal mouth opening (MIO); results showed that about 50% of participants experienced trismus—defined in the study

as an MIO of less than 36 mm. This incidence was particularly high among patients with oropharyngeal cancer, specifically, over other types of HNCs. Patients reported a range of problems, including difficulties with mouth opening, eating, drinking, dry mouth, speech disorders, and voice problems. Radiotherapy could result in edema in soft issues and dryness. Researchers in [2] emphasize the importance of speech-language pathologist evaluations to ensure vital assessments, like swallow function and safe nutrition.

One test focused on assessing the effects of concurrent chemoradiotherapy (CCRT) on voice and speech outcomes of patients with advanced HNC. In the prospective clinical trial voice and speech quality were evaluated by expert listeners using perceptual metrics and by patients themselves using a structured questionnaire. Data at three key points of time —before treatment, 10 weeks after treatment, and one year after treatment—were examined. While the study did not outline specific speech therapy interventions, it did implement preventive rehabilitation exercises for enhancing swallowing and mouth opening. Doing so highlighted the importance of including voice and speech rehabilitation in the treatment plan. Results indicate changes in voice quality post-CCRT, with perceptual evaluations indicating significant improvements or even a return to baseline at 1-year post-treatment evaluation. Many patients, however, continued to perceive their voice differently from before their illness [5].

Authors of [6] concluded that patients evaluated in a multidisciplinary clinic (e.g. oncologists, surgeons, speech-language pathologists, and other specialists) are more likely to adhere to speech–language pathology treatment recommendations, emphasizing the importance of multidisciplinary care in improving patient compliance and, potentially, optimizing outcomes for patients with HNC. Studies highlighted above suggest that the common side effect of oral cancer treatments is loss of speech and voice quality, limited mouth opening, and swallowing difficulties. Although several

rehabilitation methods are recommended, none are based on treatments benefitting from an engineering perspective. This void was filled in this work by leveraging Pixel-in-Pixel Net (PIPNet) opensource software [7, 19]—2-D facial landmark tracking algorithm—and WhisperX audio speech recognition developed by the University of Oxford [8, 18], which is an updated version of Whisper developed by OpenAI [17]. Treatment plans substantiated by automated tracking of speech quality and mouth-opening measurements that utilize image and signal processing are an intelligent strategy for advancing the field of HCN healthcare.

3 Facial Measurements

This chapter explains how facial muscle movements were tracked and measured. First, data collection setup and components used for measurement are discussed have used. Next, details are provided about software utilized to process video frames and track facial movements. Software modifications are highlighted that describe how head motion was detached from facial muscle movements, the way in which rotational head movements were eliminated, and the process for focusing on the facial frame. Displacement and distance measurement methodology within the facial frame will then be discussed detailed. Finally, information is given about the method for aligning the processed motion and audio data, as well as the technique for visualizing the audio-visual synchronized analysis.

3.1 Data Collection and Facial Tracking Model

The data collection setup was based on a synchronized audio-visual data recording made possible by a camera with dual channel microphone. The camera has a rate of 30 FPS for video recording and 44.1 KHz for audio data recording. A grid of 3x3-inch squares was set behind participants as a reference for measurements. The distance from camera to the grid was 2.5 meters. Notably, this distance can be altered, as discussed later in this chapter. During testing, participants are asked to read a passage appearing on a screen, while their facial movement and speech data are recorded. The high accuracy of deep learning models has been proven for face and facial landmark detection. Recording video data is first sliced into frames, which are each processed by a modified version of a CNN based model—FaceBoxesV2—to detect and insert a bounding box on the participant's face. This processed segment is then passed through a 2D facial landmark detection model (i.e., PIPNet) to track participant face movements. The model for each image frame detects 68 landmarks specified on eyes, eyebrows, nose, mouth, and facial border. Size and location of

bounding box changes from frame to frame due to uncertainty in face detector model and head movements. A method to compensate for this uncertainty was developed, wherein a dynamic bounding box served as a reference point and head motions were eliminated. See the next subsection for a detailed explanation.

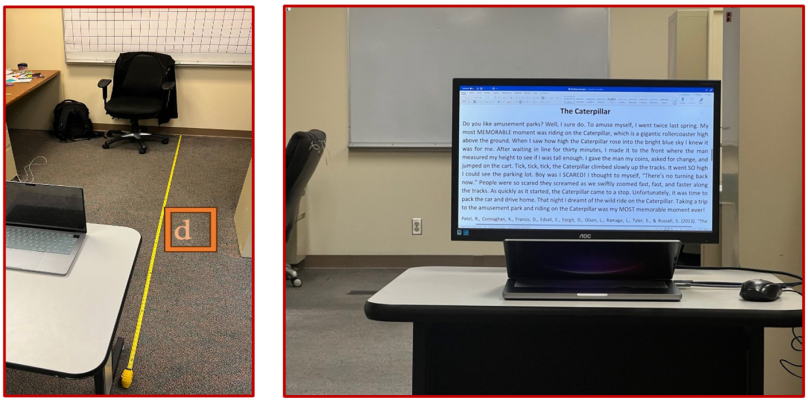

Figure 3-1 Data collection setup

3.2 Translational and Rotational Adjustments

In the original python script of PIPNet, landmarks are referenced to the top left corner of bounding boxes, which makes it difficult to detach facial motion from head movements due to changes in the location and size of the bounding box. The software script was modified to indicate the x-y location of landmarks and bounding boxes for each frame in a NumPy array. Two translations were written to detach head motion from facial muscle movements. First, original image landmarks were saved in NumPy arrays, making the top left corner of every video frame the fixed reference point. However, this fixed reference point does not eliminate head motion. Hence, landmarks were translated to the top of the participant's nose landmark. Combined, these two steps eliminated the primary head locational movement from the trajectory of landmarks, leaving only facial muscle movements, as well as rotational and up-down head motions that must be eliminated.

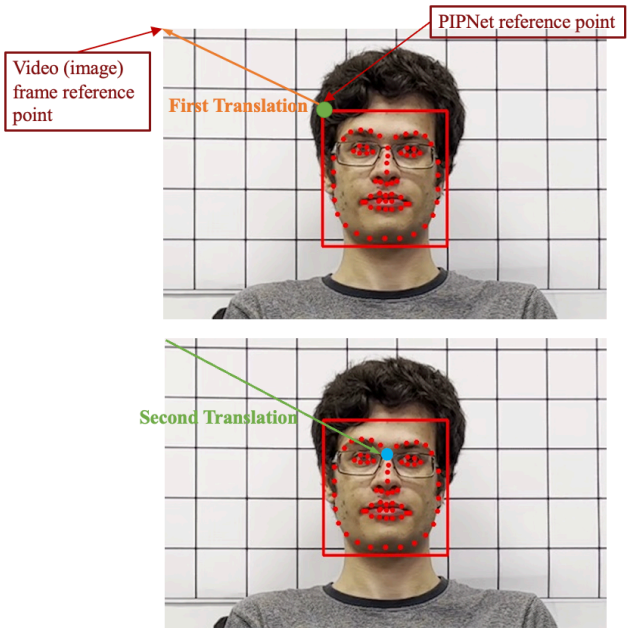

Figure 3-2 Landmarks coordinate translation.

Lateral rotation of head was measured using the angle between the middle point of the top lip relative to the y-axis. Next, landmarks were multiplied by the rotation matrix to compensate for lateral rotation.

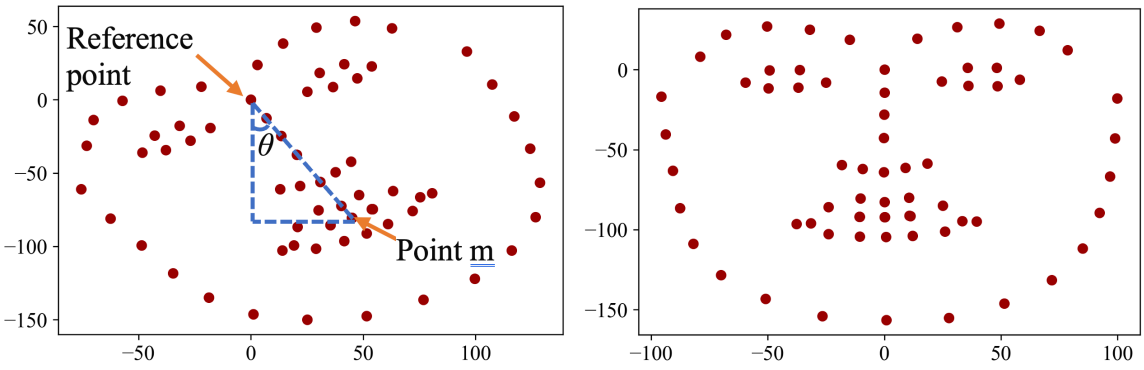

Figure 3-3 Head rotational corrections (image alignment)

The angle θ, can be calculated as [3-1].

$$\Theta = \arctan\left(\frac{x(m)}{y(m)}\right) \qquad 3\text{-}1$$

The general formula for a 2-D counterclockwise rotation is given as [3-2].

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \quad \text{3-2}$$

A clockwise rotation is required to adjust Figure 3-3 left. Given that location of point m is used for calculating θ, this point now has a negative y position, [3-1], thus a negative number for angle θ, [3-2] automatically makes a clockwise rotation, as shown in Figure 3-3.

Head up-down motion is removed using a nose length measurement. Since this factor is the most stable element among facial muscles, any change of its projection on a 2-D camera frame indicates up-down head motion. To avoid complexities caused by increasing the number of cameras in the recording setup and becoming aware that this motion is inconsequential during the speech task, a 2-D up-down rotation correction equaling a nose length-based scaling factor was multiplied by the location of all translated landmarks to maintain a consistent length. Eventually, only facial frame muscle movements remain, which accurately provide displacements.

3.3 Measurements

To measure either landmark displacement or the distance between landmarks, a projected length of the object was required for the camera sensor. Here we provide an example case of the measurement methodology. Calculating pixel amount for the 3*3 inch (76.2*76.2 mm) marked reference grid was necessary. Distance between face surface (i.e., facial frame) and background grid suggested scaling a correction factor in pixels relative to the grid on the facial frame. The original size of the grid at distance of 2.5 meters from the camera measured 51*51 pixels. The distance of face from the grid was approximately 1 Ft (= 0.3048 m). The scaling correction factor for grid size with 2.5 meters distance from the camera to grid is shown in [3-3].

$$\text{Scaling correction factor} = \frac{\text{Distance of camera from grid}}{\text{Distance of camera from grid} - \text{Distance of grid from face}}$$

$$= \frac{2.5}{2.5 - 0.3048} = 1.1388 \quad \text{3-3}$$

Hence, corrected pixel size of the grid is 1.1388*51 or 58 pixels. Camera focal length was 26 mm. Length of reference grid in the camera sensor is given in formula [3-4].

$$\text{Distance of camera to object (mm)} = \frac{\text{object length in real world (mm)} * \text{focal length (mm)}}{\text{object length in camera sensor (mm)}} \quad 3\text{-}4$$

$$2500\ (\text{mm}) - 304.8\ (\text{mm}) = \frac{76.2\ (\text{mm}) * 26\ (\text{mm})}{\text{x (mm)}} \rightarrow \text{x} = 0.9025\ (\text{mm}) \quad 3\text{-}5$$

Knowing the length of reference grid in the camera sensor (0.9025 mm) provides a proportional relationship to find the length of a referenced 63.5 mm object in a facial frame with 49 pixels [3-6].

$$\text{y (mm)} = \frac{49\ (\text{pixels}) * 0.9025\ (\text{mm})}{58\ (\text{pixels})} = 0.76\ (mm) \quad 3\text{-}6$$

Accordingly. object length for the camera sensor can be calculated for the real world [3-7] with indicated error [3-8].

$$2500\ (\text{mm}) - 304.8\ (\text{mm}) = \frac{\text{L (mm)} * 26\ (\text{mm})}{0.76\ (\text{mm})} \rightarrow \text{L} = 64.45\ (\text{mm}) \quad 3\text{-}7$$

$$\text{error} = \left| \frac{64.45 - 63.5}{63.5} \right| = \%\ 1.496 \quad 3\text{-}8$$

A summary of measured object lengths at various camera-to-grid distances between 2.5 and 4 meters is provided in Table 3-1.

Table 3-1 Measurement errors at different distances of camera to the reference grid and object

Distance from camera to grid (meter)	Object Length (pixels)	Calculated Length (mm)	Error
2.5	49	64.45	% 1.496
3	39	63.81	% 0.852
3.5	32	62.11	% 1.496
4	28	62.27	% 1.936

3.4 Audio and Motion Alignment

To visualize lip movements during the pronunciation of specific words (e.g. “caterpillar”), time intervals of the specific word and segmented audio/motion array data is required. Frame-by-frame lip movement progression had to be saved as .jpeg files and attached to segmented audio and .jpeg files to align motion and voice data. Since video frame rate (i.e., 30 FPS) is considerably less than audio recording frequency (i.e., 44.1 KHz), a maximum miss alignment error of 1/30 (or 33 milliseconds) was obligatory. Figure 3-4 visualizes the motion of four points (e.g., top, down, left and right locations of lip) over time for pronouncing the word “caterpillar”. The bottom right subplot in Figure 3-4 illustrates the progression of lip horizontal axis (x) position over time. This plot clearly demonstrates the accurate translational and rotational adjustments impact discussed earlier. The upper and lower lip landmarks are positioned on the origin of horizontal axis (or x-axis) over time, since they are always aligned with the nose for a healthy person within the facial frame. Furthermore, left and right lip corners have a symmetrical trajectory along this axis for a

healthy person, which is also demonstrated in Figure 3-4. Results indicate the power of the head motion corrections that are made possible by this research. The shrinking effect of corners during pronunciations indicate that the effect is a personalized characteristic and can be different from person to person. Compare motion in Figure 3-4 with Figure 3-5. The top right plot shows the progression of lip vertical (y) position over time, clearly demonstrating mouth opening and closing over time wherein the upper lip (blue) and lower lip (orange) have a larger and smaller gap between each other. For example, the circled region shows where the sound /p/ happens, and the time interval (0 sec – 0.15 sec) demonstrates the mouth opening for pronouncing /ca/ phoneme.

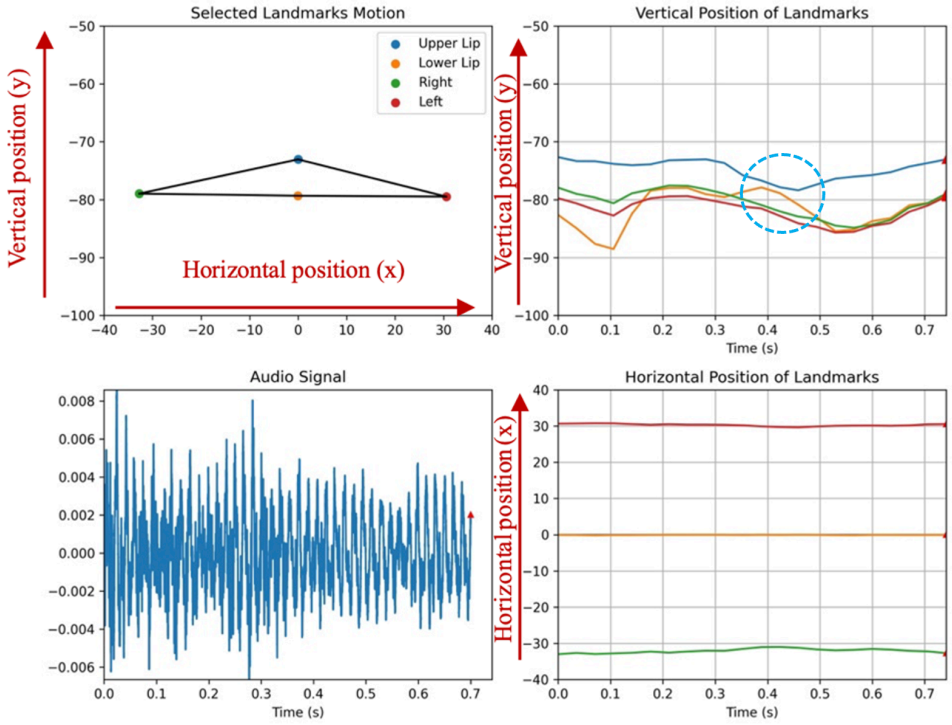

Figure 3-4 First subject's lip motion pronouncing the word "caterpillar".

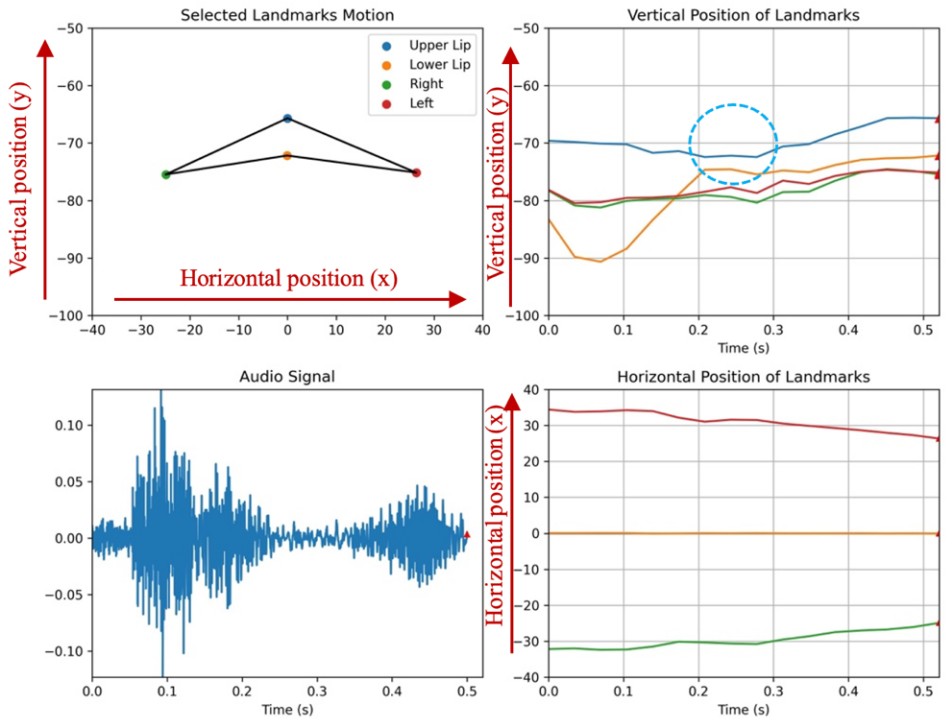

Figure 3-5 second subject's lip motion pronouncing the word "caterpillar".

3.5 Motion Features

Three distances were selected to obtain the oral cavity related measurements for a motion-and-audio feature space after each recording. These measurement features are visualized in Figure 3-6. Maximum vertical distance is the maximum distance between the upper and lower lip while pronouncing a single word. Maximum horizontal distance and minimum horizontal distance are the maximum and minimum width of the lips, respectively, while pronouncing a word. A time segment was derived for each word which will be explained in the voice analysis chapter. A visualization of these distances is shown in Figure 3-6. After measuring the values in pixels and depending on the distance between camera and subject, lip distances were calculated in millimeters, utilizing the method explained in the measurements section, above. The features were saved in the feature space table for each participant. This information provides a summary of the

amount of mouth opening and shrinking effects on sides of the lips and helps track changes resulting from either pain or speech therapy following oral cancer surgery. Table 3-2 offers examples of abovementioned features for each word in “The caterpillar” passage.

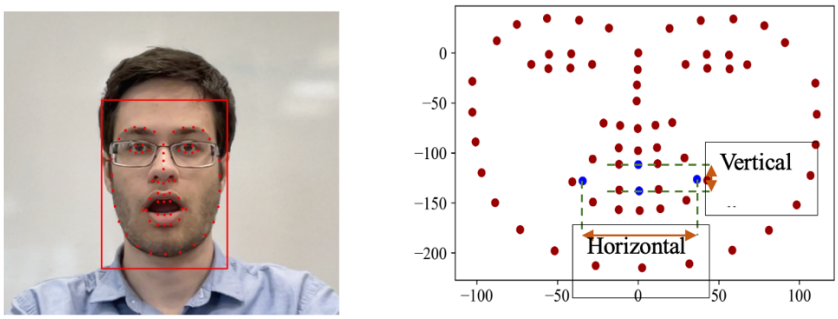

Figure 3-6 Selected motional distance features for the feature space.

Table 3-2 Lip Motion Measurements

word	max_vertical_distance (mm)	max_distance - horizontal(mm)	min_distance - horizontal(mm)
do	8.89	47.06	43.20
you	11.48	44.16	41.29
like	21.76	59.44	48.30
amusement	6.93	55.87	49.91
parks	9.01	53.96	47.87
well	20.93	52.28	40.68
i	17.30	57.11	55.07
sure	9.63	51.93	44.56
do	7.57	48.04	40.99

3.6 Lip Displacements Distributions

Insights from the comparison of the distribution of lip displacement of normal and the emulated case study is anticipated to be valuable for healthcare providers to design speech therapy and rehabilitation protocols, especially those focusing on exercises for enhancing articulation. Analyzing the ways in which artificial constraints affect speech can provide clues about the mechanics of various speech impediments, as well as guide diagnosis and intervention strategies.

When compared with the baseline normal condition, two simulations (e.g., keeping an oral obstacle under the tongue and clenching teeth while reading a passage "the caterpillar") demonstrated two behaviors to compensate for limitations resulting from tongue position and jaw displacement. Measurements of simulated conditions from the first participant (see first row of histograms in Figure 3-7) shows nearly the same behavior for maximum lip distance between the upper and lower lips. The third column shows only a slight change in skewness. However, measurements of simulated conditions for the second participant (see second row of histograms in Figure 3-7) show lower mean and standard deviation for maximum vertical distances when teeth were clenched (see third column) when compared to normal teeth positioning (see first column). These results indicate that both conditions pose potential challenges to clear articulation, as they mimic particular speech impediments.

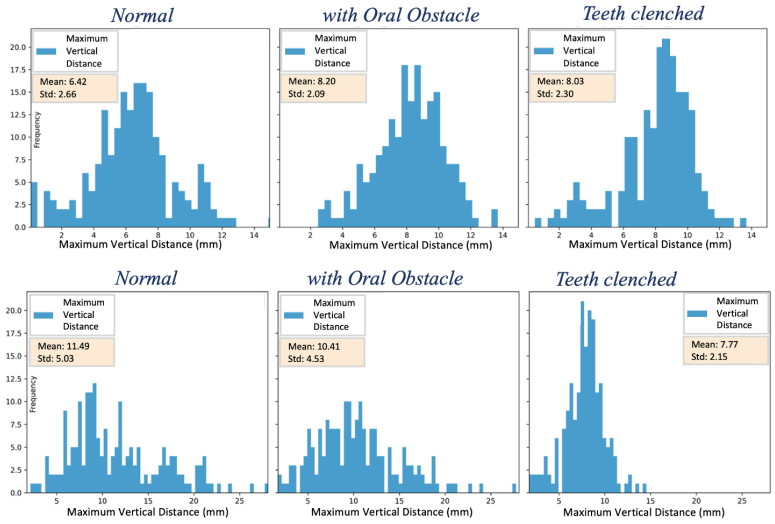

Figure 3-7 Distribution of motional feature "maximum vertical distance" (maximum distance between the upper lip and the lower lip) while reading "the caterpillar" passage under normal condition and emulated by oral obstacle under the tongue and teeth clenched together. First participant on the first row and second participant on the second row.

Different motional behavior is also evident from the maximum and minimum width of the lips while reading "the caterpillar" passage. Measurements for the first subject (see first row of Figure 3-8) under normal condition reveal a more concentrated distribution for maximum and minimum

width. However, under simulated conditions, the spread of values (i.e., standard deviation) increases significantly with a nearly 78% and 73% increase in maximum horizontal distance (see Figure 3-8) and minimum horizontal distance (see Figure 3-9), respectively. On the other hand, measurements for the second subject show no significant change for minimum and maximum lip width distribution.

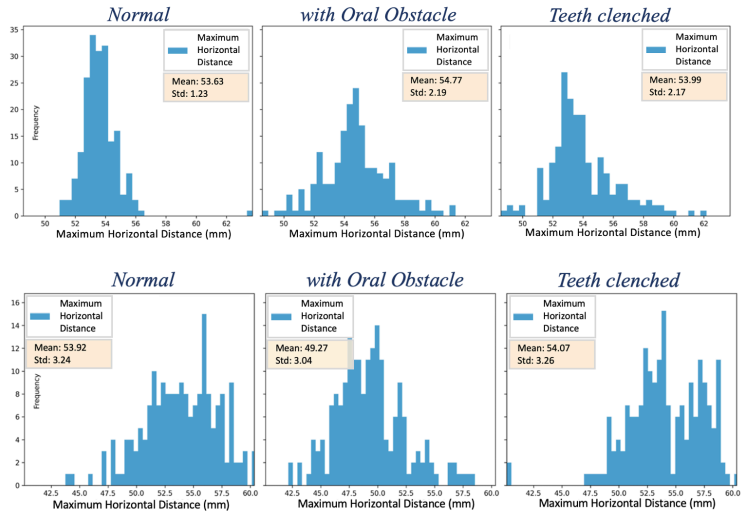

Figure 3-8 Distribution of motional feature "maximum horizontal distance" (maximum width of the lips) while reading "the caterpillar" passage under normal condition and emulated by oral obstacle under the tongue and teeth clenched together. First participant on the first row and second participant on the second row.

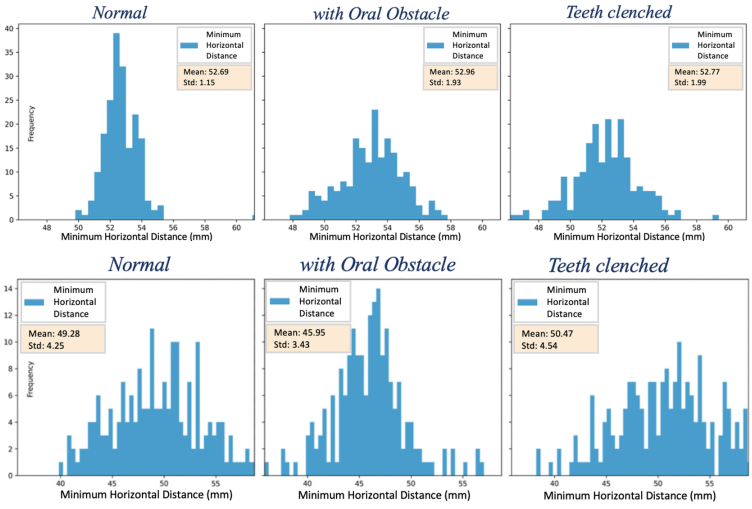

Figure 3-9 Distribution of motional feature "minimum horizontal distance" (minimum width of the lips) while reading "the caterpillar" passage under normal condition and emulated by oral obstacle under the tongue and teeth clenched together. First participant on the first row and second participant on the second row.

3.7 Summary

This chapter provided an in-depth exploration of methods used to track and measure facial muscle movements: data collection setup, adjustments for head and facial movements, displacement and distance measurement methodology, and synchronizing motion with audio for analysis. The setup involved synchronized audio-visual recordings using a camera and dual-channel microphone, with a specific protocol to facilitate precise measurements with a reference grid. Subject data, including the capture of facial movements and speech sound, were recorded foreach while he/she read a passage “the caterpillar.”, Utilizing deep learning models (e.g., FaceBoxesV2 [13, 14] and PIPNet [7]), faces and facial landmarks across frames were detected, which enabled detailed tracking of facial muscle movements. Modifications were made to PIPNet’s original script to accurately separate facial movements from head motions. Doing so involved translating landmark positions to a fixed reference in the facial frame and compensating for head movements, including both translational and rotational adjustments that were based on landmark positions. These adjustments enabled the isolation of pure facial muscle movements from other head movements. The process for synchronizing audio and motion data explained how the visualization of lip movements was aligned with spoken words. This chapter also demonstrated the effectiveness of the earlier described adjustments necessary for accurately tracking and visualizing facial movements to synchronize with audio data. Finally, the way in which oral-related measurements were derived from recorded data—including distances between key points on the lips during speech—was provided. These measurements are critical for analyzing speech dynamics and can be used to track changes over time, especially those resulting from medical treatments or therapy. Overall, this chapter outlined a comprehensive approach for accurate measurement and analysis of facial

movements, providing a framework for facial measurements, which is necessary for analyses detailed in upcoming chapters.

4 Voice Analysis

This chapter details speech audio signal analysis and the importance of each audio feature in the study methodology. First is a brief overview of the audio recognition software utilized to segment audio files into individual words that are detected at specific time intervals. Next, audio features (e.g., time duration, RMS sum, zero crossings, pitch [or fundamental frequency], and formant frequencies) are defined, and the way in which each was extracted for the model is described. Special attention is given to formant frequency distributions derived by signal processing methods to provide a foundation for the comparison between normal and restrictive conditions and to characterize the loss, gain, and shift of frequencies.

4.1 Speech Audio Segmentation

Start and end times for the utterance of each word is necessary to extract audio features. To extract these time intervals, detached audio files of recordings were put into the speech recognition software. In this way, speech characteristics per pronounced words can be measured for analyses.

Table 4-1 Audio Segmentation Examples

word	Start time	End time	Duration (sec)
do	1.668	2.008	0.3400000000000000
you	2.349	3.389	1.0400000000000000
like	4.829	5.029	0.2000000000000000
amusement	5.129	5.629	0.5
parks	5.669	5.969	0.3000000000000010
well	6.329	6.569	0.2400000000000000
i	6.769	6.849	0.08
sure	6.869	7.07	0.2010000000000010
do	7.09	7.29	0.2000000000000000

Understanding the behavior of the speech signal is necessary for characterizing audio features. During pronunciation of a word, a speech signal is defined as the combined effect of the vibration of vocal cords and the shape of the vocal tract. The latter is important because it determines how

air passes through the mouth, which effects pronunciation. When examining the speech signal for the word “scared”, for example, (see Figure 4-1), phonemes can be determined by looking only at the periodicity of the signal, as explained below. The zero-crossing rate (ZCR) for the unvoiced consonant /s/ is much higher when compared to other consonants and vowels. In fact, there is a high frequency content carried by the signal in this part of the word. Periodicity also changes for different phonemes when examining the signal for consonants /c/, /a/, /r/, and /ed/. Figure 4-1 demonstrates the importance of frequency-related features (e.g., ZCR, pitch, and formants) to be analyzed under both regular and emulated conditions.

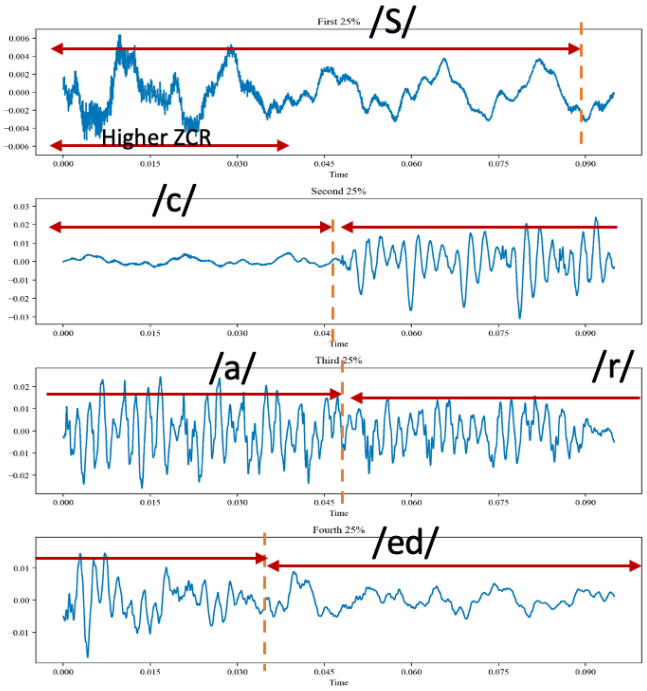

Figure 4-1 Raw Audio Speech Signal Behavior for Different Phonemes.

4.2 Audio Features

4.2.1 Time Duration

One simple, yet principal feature for analysis is the time duration (i.e., length) of the pronunciation of a word. It is important to note that this can change as a result of loss of vocal tract components after oral surgery.

4.2.2 RMS Sum

A more complex feature utilized for analysis of this same variable is the root mean square (RMS) of an audio signal. The RMS value of an oscillating AC signal represents its DC component. For this study, the RMS value represents the loudness of the audio [4-1]. The RMS was calculated for small time frames (e.g., 20 ms), and then summed, to stay within the time interval of each spoken word.

$$RMS = \sqrt{\frac{1}{n}\sum_{i} x_i^2} \qquad 4\text{-}1$$

$$RMS\_sum = \sum_{j} \sqrt{\frac{1}{n}\sum_{i} x_i^2} \qquad 4\text{-}2$$

where j is the number of frames in the equation 4-2.

4.2.3 Zero Crossing Rate

A common way to measure the smoothness of a signal is counting the number of zero crossings within a time interval [9]. Voiced signals (e.g. /z/, /v/) are produced by the vibration of the vocal cords in addition to the way in which the vocal tract is shaped. Unvoiced signals (e.g. /s/, /f/) do not require vocal cord vibration. Typically, voiced signals have lower zero crossings than unvoiced signals. This specification can be affected by modifications in the vocal tract shape, which likely

occurs as a consequence of oral surgery. The plots shown in Figure 4-2 through Figure 4-5 serve as examples of measured ZCR for a specific sentence aligned with the audio amplitude. The figures illustrate high ZCR for unvoiced signal segments that were measured for two participants under normal conditions (see Figure 4-2 and Figure 4-3). The start and end of each word is indicated by dashed lines, and the word being pronounced is positioned between those dashed lines. It is worth mentioning that ZCR varies across individuals. The plots for phonemes /d/ and /s/ show that the second participant has a much higher ZCR because of dissimilar accent and frequency content. When listening to the audio signal for that participant pronouncing the word "do," a partial sound of /z/ is audible in the pronunciation.

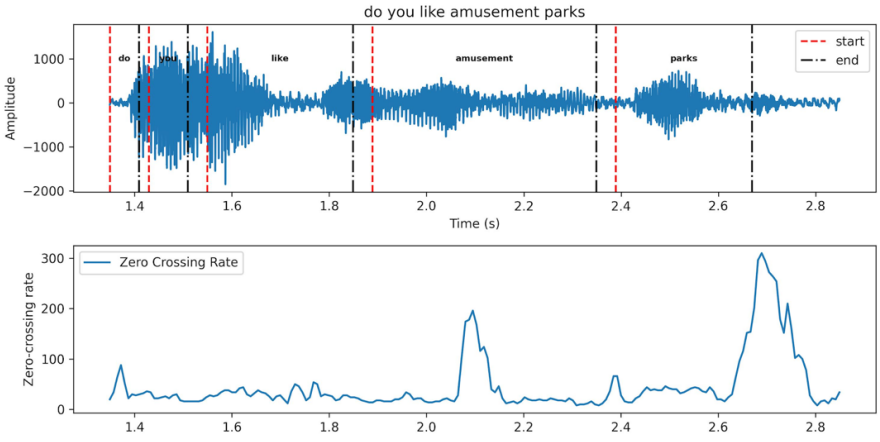

Figure 4-2 First Subject under Normal Condition

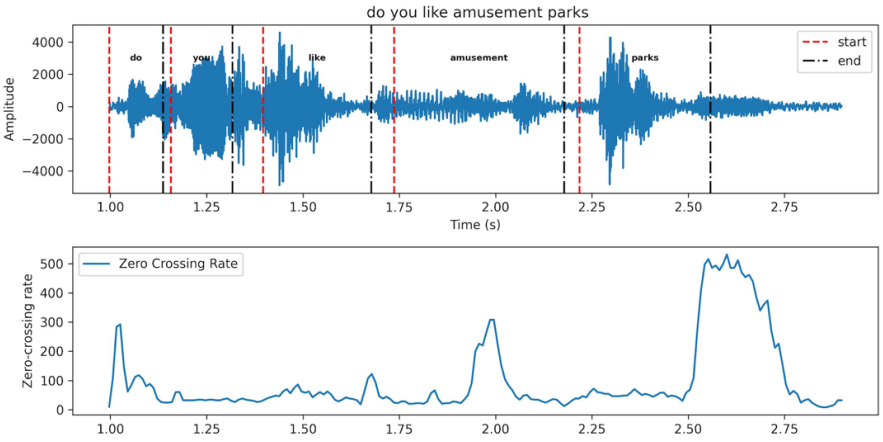

Figure 4-3 Second Subject Normal Condition

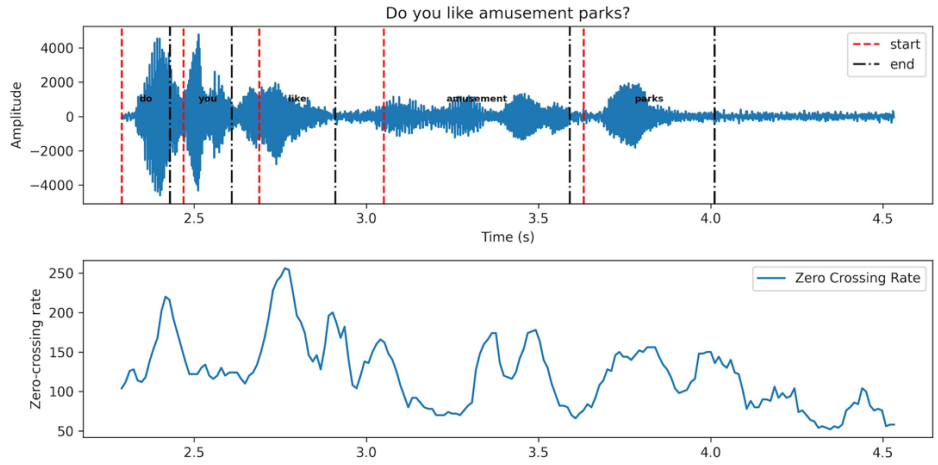

Figure 4-4 First Subject Emulated with Oral Obstacle Under the Tongue

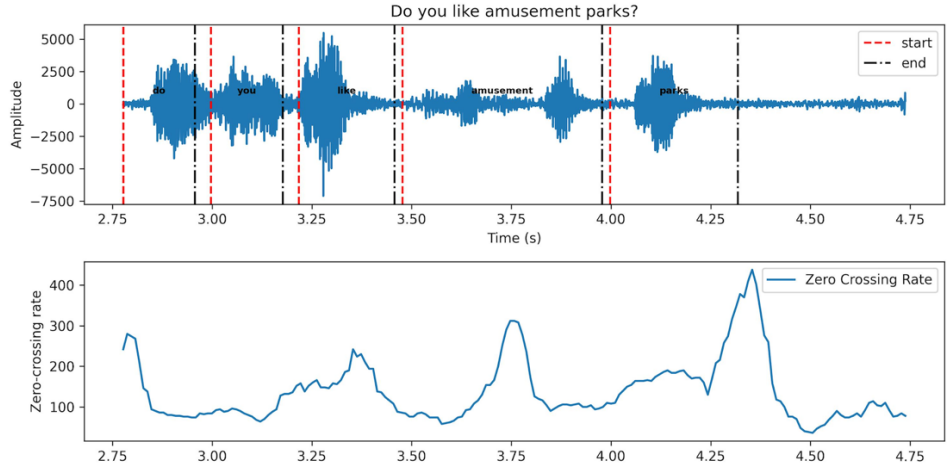

Figure 4-5 Second Subject Emulated with Oral Obstacle Under the Tongue

When comparing the ZCR produced under regular and emulated conditions, it is apparent that the overall rate increased for the given sentence, even though the peak is lower. This phenomenon is a consequence of tongue positioning, which limits the tongue's motion and keeps its front component on a lower position and its back component on a higher position. This positioning makes it difficult to pronounce unvoiced sounds (e.g. /s/) clearly and decreases the vocal tract tube volume. Figure 4-6 and Figure 4-7 show that the distribution's variability (e.g. std) of ZCR for "the caterpillar" passage words is decreased for both participants under emulated conditions. The first participant has a standard deviation of 2191 under normal conditions and 1285 and 1239 for

emulated conditions, as shown in Figure 4-6. Standard deviation for the second participant is 3222 under normal conditions and 1627 and 2585 for the emulated conditions, as shown in Figure 4-7.

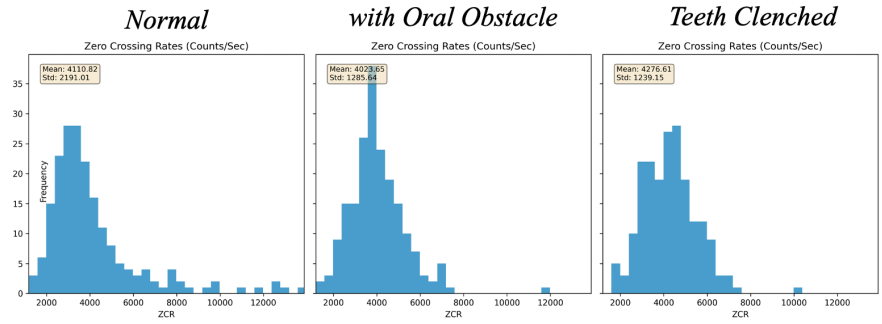

Figure 4-6 From left to right respectively: First subject's ZCR per word distribution while reading the caterpillar passage under normal (i.e., regular) condition; reading with oral obstacle; and reading while teeth are clenched together.

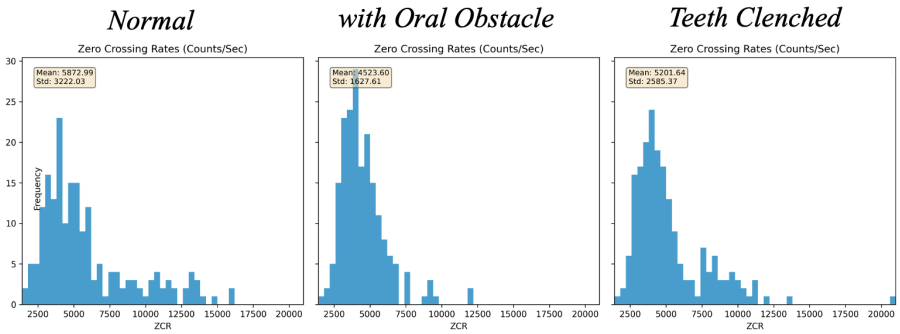

Figure 4-7 From left to right respectively: Second subject's ZCR per word distribution while reading "the caterpillar" passage under normal (i.e., regular) condition; reading with oral obstacle; and reading while teeth are clenched together.

4.2.4 Pitch (Fundamental Frequency)

Pitch frequency is directly related to oscillations originating from vocal cords. Fundamental frequency (F0) typically fluctuates throughout a sentence —rather than remaining static—due to the variability in speech signal and pronunciation of various vowels. MATLAB's built-in "pitch" function was used to estimate the F0 for a word in a ".wav" file. A histogram was derived for each word, and the F0 with the highest count was selected as the dominant pitch frequency of the pronounced word for the feature space. The bin size in the histogram was set at 4.5 Hz. The plots shown below visualize pitch frequencies for the word "memorable" under normal condition and when emulated with an obstacle under the tongue.

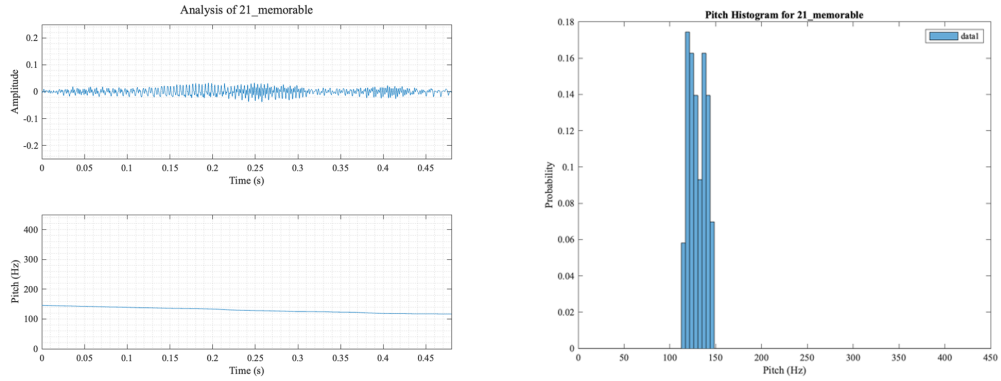

Figure 4-8 First participant's pitch (i.e., fundamental frequency) while pronouncing the word "memorable" under normal condition.

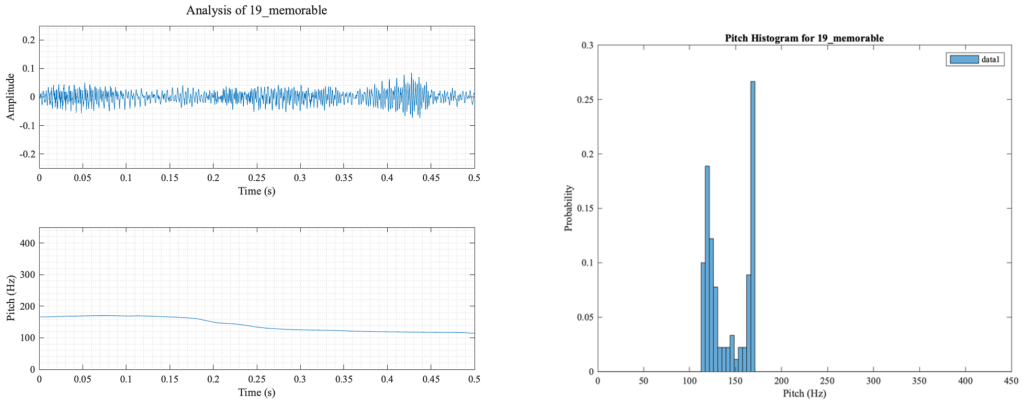

Figure 4-9 First participant's pitch (i.e., fundamental frequency) while pronouncing the word "memorable" under emulated condition with oral obstacle under the tongue.

4.2.5 Formant Frequencies

Peaks in the envelop of the audio spectrum represent high energy areas of the spectrum. These frequencies are known as formants, and each corresponds to a resonance in the vocal tract. The vocal tract is tube shaped and closed at one end by vocal folds and at the other by the lips. The shape of the vocal tract's cross-sectional area is modulated by the positioning of the tongue, lips, jaw, and soft palate (i.e., velum). The resulting spectrum of the vocal tract's response includes a series of resonance frequencies (i.e., formants) that are unique to the tract's shape. For a short time-frame, formant's location specifies the vowel that was pronounced [10, 11]. The location of these frequencies is closely related to the shape of vocal tract. Pain or loss of muscles that shape the

vocal tract and/or muscle movements often change after oral cancer surgery and, subsequently, change formant's location. Thus, analyzing these frequencies is crucial for tracking the quality of speech after surgery. A common and well-known method for extracting the frequency envelop in speech processing is Linear Predictive Coding (LPC).

4.2.5.1 Linear Predictive Coding

The hypothesis underlying linear predictive analysis is that a speech sample can be estimated from a linear combination of past speech samples. Unique predictor coefficients are determined by minimizing the squared error between predicted and actual speech samples [11].

Speech samples are related to the excitation pulse as expressed in equation 4-3:

$$s(n) = \sum_{k=1}^{p} \mathrm{a}_k \, s(n-k) + Gu(n) \qquad 4\text{-}3$$

Between pitch pulses Gu(n) is assumed to equal zero, so that predicted s(n) is a linear weighted discrete sum over past speech samples. Given that Gu(n) is nonzero, s(n) can be estimated approximately.

Let us assume speech signals are obtained via a linear predictor with α_k coefficients. Output is calculated, as follows.

$$\hat{s}(n) = \sum_{k=1}^{p} \alpha_k \, s(n-k) \qquad 4\text{-}4$$

This equation indicates that the n^{th} sample can be predicted from last p samples. Error between the actual and predicted speech signal is shown in equation 4-5.

$$e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{k=1}^{p} \alpha_k \, s(n-k) \qquad 4\text{-}5$$

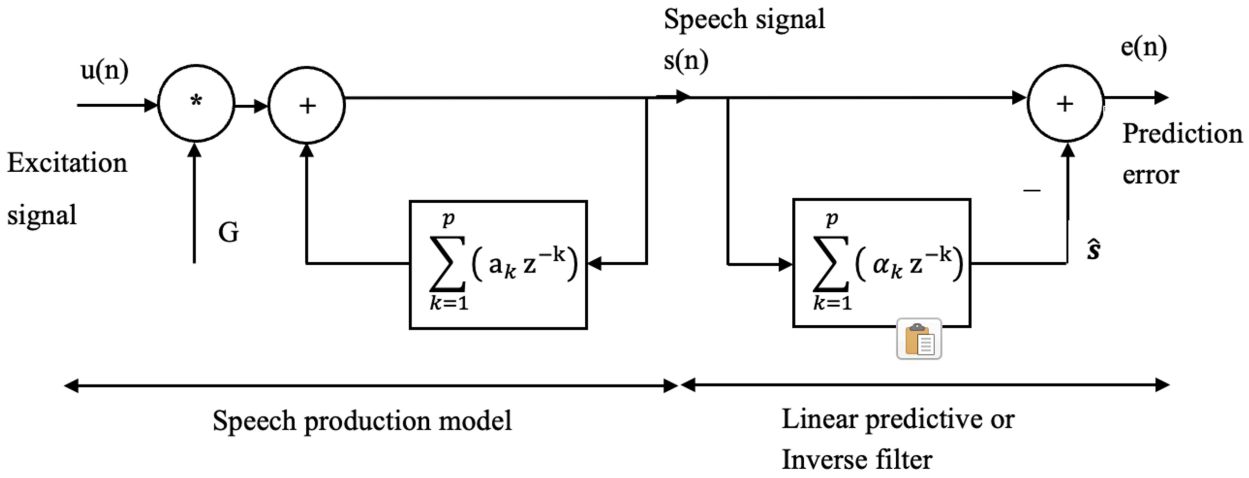

Figure 4-10 Linear prediction model block diagram.

4.2.5.2 Estimation of the Prediction Coefficients

To minimize the squared error and find the coefficients, a condition must be found that satisfies the derivative of squared error equal to 0:

$$\frac{\partial E}{\partial \alpha_i} = 0 \qquad for \quad i = 0, 1, \dots, p \ (usually\ 10\ to\ 14) \tag{4-6}$$

$$\frac{\partial E}{\partial \alpha_i} = \sum_{n=1}^{N} 2\,[s(n) - \sum_{k=1}^{p} \alpha_k\, s(n-k)]\,[-s(n-i)] = 0 \tag{4-7}$$

$$\sum_{n=1}^{N} s(n)\, s(n-i) - \sum_{n=1}^{N} \sum_{k=1}^{p} \alpha_k\, s(n-k)\, s(n-i) = 0 \tag{4-8}$$

$$\sum_{n=1}^{N} s(n)\, s(n-i) - \sum_{k=1}^{p} \alpha_k \sum_{n=1}^{N} s(n-k)\, s(n-i) = 0 \quad for \quad i = 1, 2, \dots, p \tag{4-9}$$

$$\sum_{k=1}^{p} \alpha_k \sum_{n=1}^{N} s(n-k)\, s(n-i) = \sum_{n=1}^{N} s(n)\, s(n-i) \quad for \ \ i = 1, 2, \dots, p \tag{4-10}$$

The definition of the autocorrelation function can be found using equation 4-11.

$$R(k) = \sum_{n=0}^{N} s(n)\, s(n+k) \tag{4-11}$$

The left side of the previous equation can be replace n-k by l and obtain 4-12.

$$\sum_{l=1}^{N} s(l)\, s(l+k-i) = R(k-i) \tag{4-12}$$

Note that in autocorrelation, R(k) = R(-k) for the right side of 4-10 to obtain 4-13.

$$\sum_{n=1}^{N} s(n)\, s(n-i) = R(i) \tag{4-13}$$

Subsequently, by substituting 4-12 and 4-13 in 4-10, we obtain:

$$\sum_{k=1}^{p} \alpha_k\, R(k-i) = R(i) \quad for \quad i = 1, 2, \dots, p \tag{4-14}$$

This equation can be written in matrix form 4-15:

$$\begin{pmatrix} R(0) & R(1) & \cdot & R(p-2) & R(p-1) \\ R(1) & R(0) & \cdot & . & R(p-2) \\ . & R(1) & \cdot & . & . \\ R(p-2) & . & \cdot & R(0) & R(1) \\ R(p-1) & R(p-2) & \cdot & R(1) & R(0) \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ . \\ \alpha_{p-1} \\ \alpha_p \end{pmatrix} = \begin{pmatrix} R(1) \\ R(2) \\ . \\ R(p-1) \\ R(p) \end{pmatrix} \tag{4-15}$$

To obtain coefficients, we need to calculate R^{-1}, then:

$$\underline{\alpha} = R^{-1}\underline{r} \tag{4-16}$$

R, known as the Toeplitz matrix, is symmetric, and elements on the main diagonal are always equal. This equation can be solved using iterative methods, such as Durbin's algorithm. After determining LPC coefficients, the frequency response of the obtained filter from equation 4-4 provides the spectral envelope, which represents the smooth curve capturing harmonics peaks or signal spectrum formants. Formants are resonant frequencies of the vocal tract and are crucial in determining the characteristics of speech sounds. LPC coefficients model the vocal tract as a series

of filters, and thus, the spectral envelope derived from LPC coefficients reflect the formant structure of the speech signal. Furthermore, the relationship between LPC coefficients and the spectral envelope can be understood through the filter model used in LPC. The signal is modeled as being produced by exciting a linear predictive filter with a specific set of LPC coefficients. The filter's frequency response characterizes the spectral envelope.

4.2.5.3 Formant Frequency Determination

Five formant frequencies were derived in this study using LPC for each word in two ways. First, to intuitively find an averaged frequency behavior of a word and to have the same number of features for each word, a time window for LPC was selected as the same time duration of each word and extracted a unique number of formants per word for the feature space. This reveales the average of formants across different vowels and phonemes within the word. Corresponding spectral envelop and formants for the sample word “caterpillar” are visualized in Figure 4-11 for two participants under both regular and emulated conditions for five trials. Minor differences are evident when comparing the spectral envelopes. These are primarily due to the context and pronunciation; however, a similar pattern is repeated throughout.

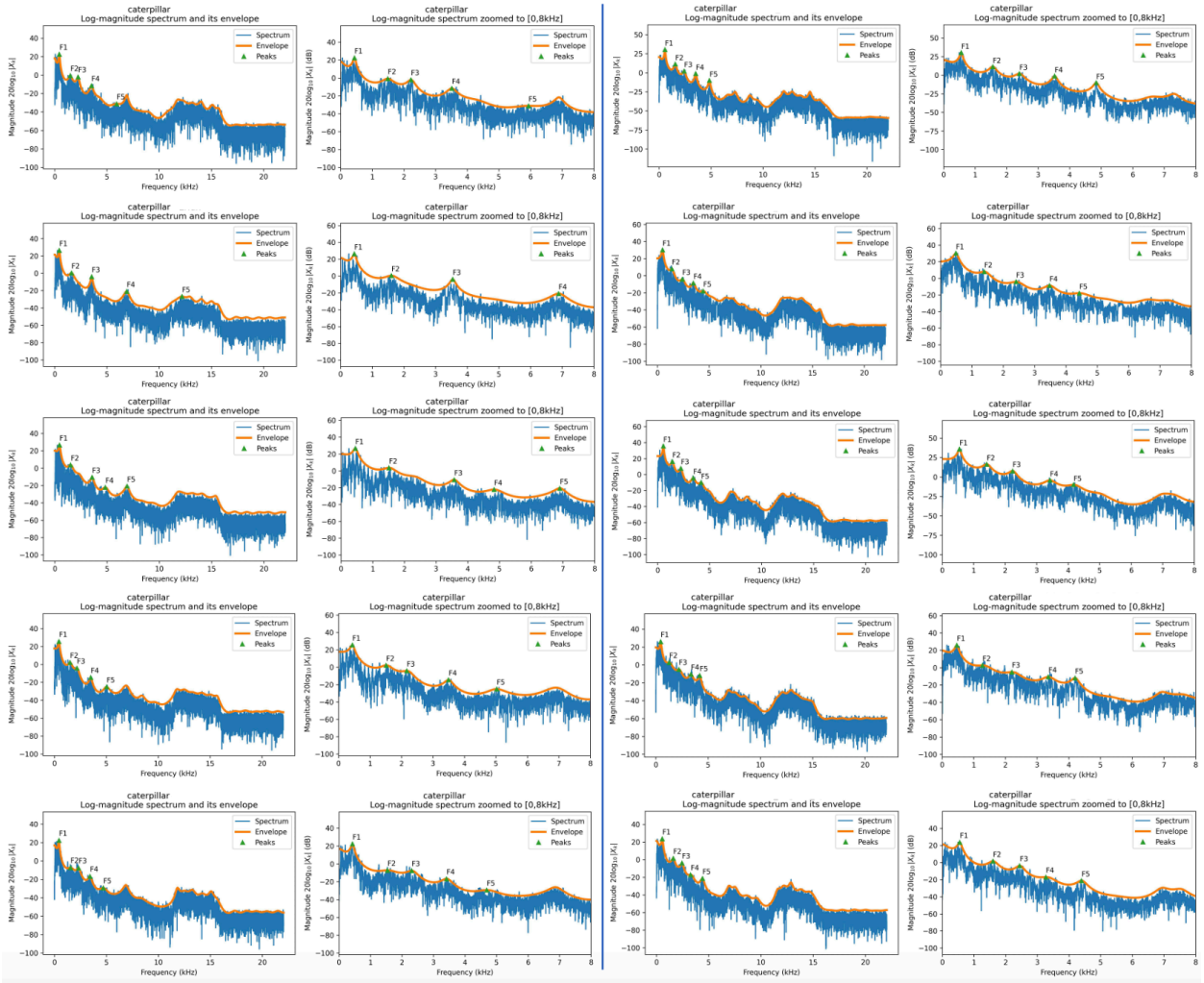

Figure 4-11 Spectral envelopes and formant frequencies for five trials of the word "caterpillar" under normal condition pronounced by two participants: First participant is indicated in Column1, and the second in Column2. The Second Subplot on each row of columns 1 and 2 is has been magnified from 0 to 8 KHz.

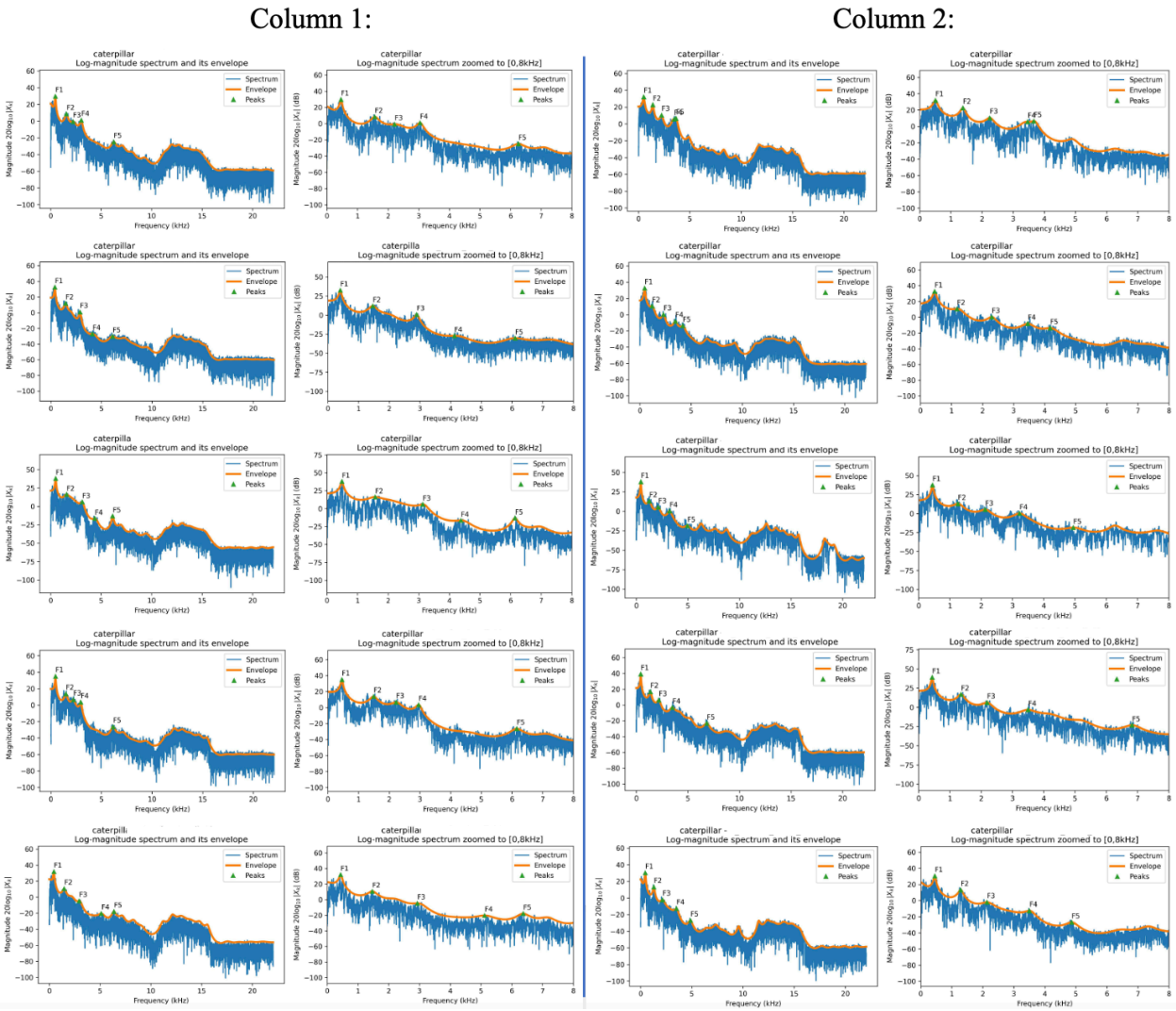

Figure 4-12 Spectral envelopes and formant frequencies for five trials of the word "caterpillar" under an emulated condition with an oral obstacle under the tongue; pronounced by two participants; first participant in Column1 and second participant in Column2. Second subplot on each row of Column1 & 2 is magnified from 0 to 8 KHz.

Second, oral cancer surgery can cause the loss of ability to preserve formant frequencies or can lead to a shift in frequencies and the appearance of new frequencies. To analyze these changes, formants with high temporal resolution are required. Consequently, formants with 25 milli-second time frames were derived within the time interval for words under investigation. 25 milli-seconds was the optimal window size, because the speech signal is nearly stable at this point and behavior is constant. Smaller window size values resulted in frequency resolution loss; larger window size

resulted in loss of temporal resolution. The right side of following plots shows a histogram with five formants for all word frequencies in “the caterpillar” passage that were combined under both regular and emulated conditions. The left side shows differences after subtracting five frequency formants from an emulation under normal conditions. Figure 4-13 illustrates a shift of frequencies in formants 4 and 5.

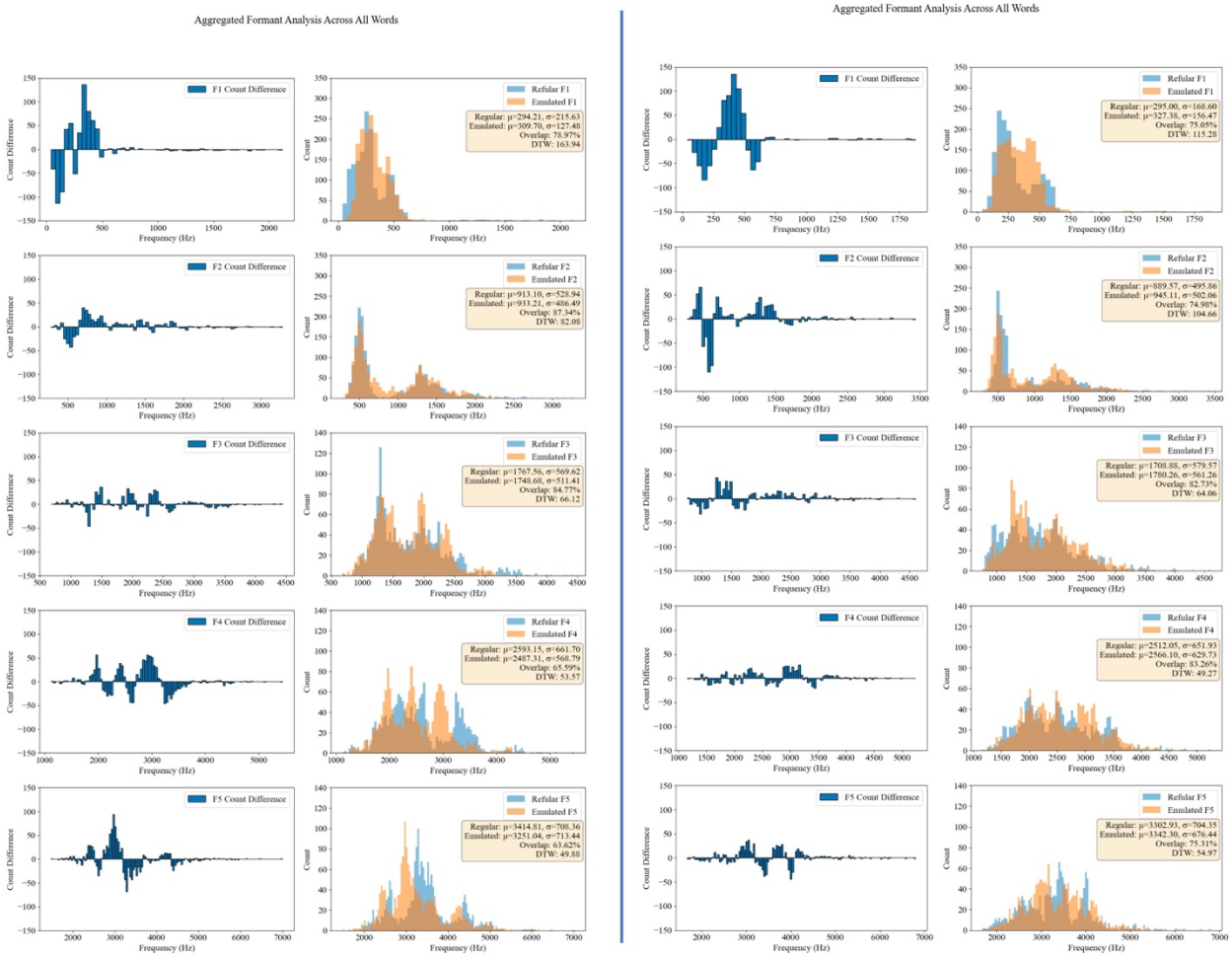

Figure 4-13 Distribution of the five formant frequencies and the subtraction of normal from emulated condition (with oral obstacle) for the first participant in the left columns and for the second participant in the right columns.

Dynamic Time Warping (DTW) and cross-correlation were calculated to investigate similarities between normal and emulated distributions and to determine frequency shift between them, respectively. DTW is an algorithm used for measuring similarity between two temporal sequences

that might vary in speed, cross-correlation is a method used to determine shift between two or more time series, wherein maximum value indicates the amount of shift DTW values have specific, converse interpretations: low DTW indicates a high degree of similarity between sequences, and high DTW value indicates a low degree of similarity.

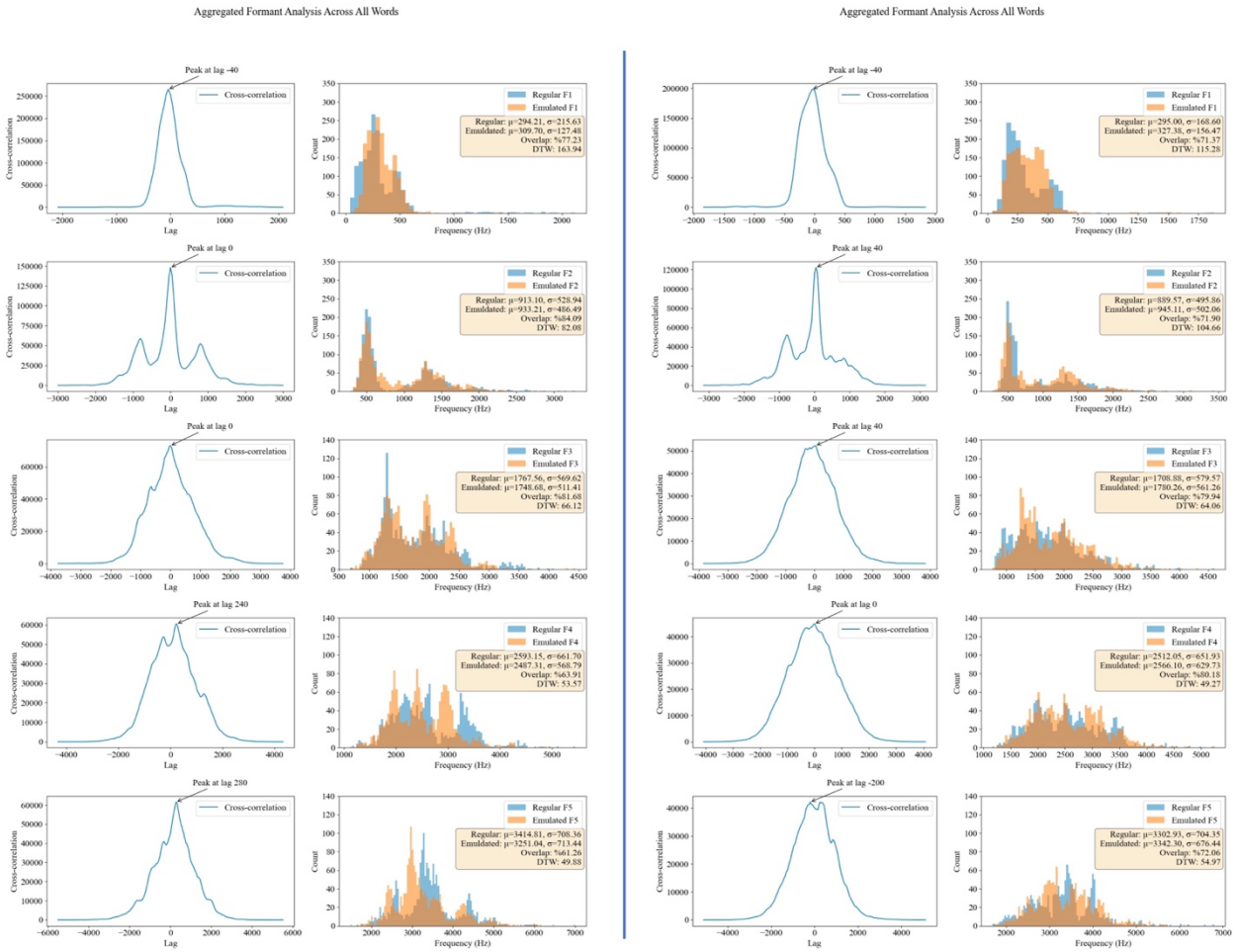

Figure 4-14 Distribution of the five formant frequencies and cross correlation between normal and emulated conditions(with oral obstacle) for the first participant in left columns and for the second participant in the right columns.

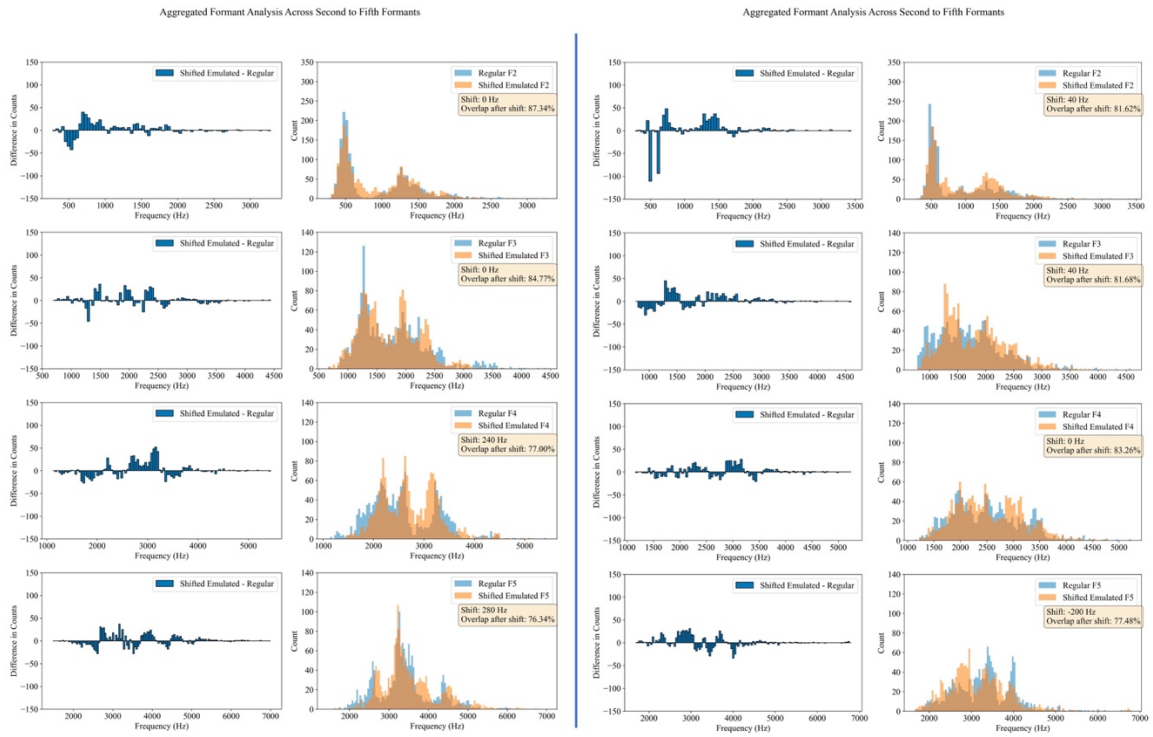

Figure 4-15 Distribution of the five formant frequencies after the shift of emulated condition and the subtraction of normal from emulated condition for the first participant in the left columns and for the second participant in the right columns.

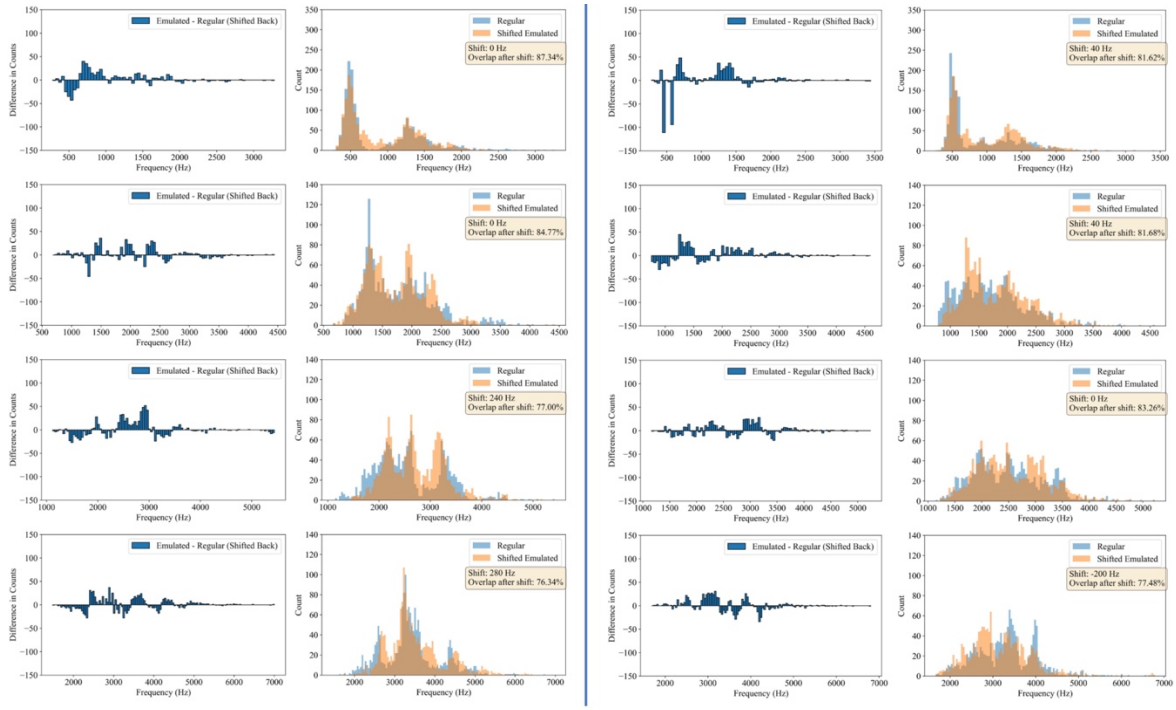

Figure 4-16 Distribution of the five formant frequencies after shifting back of emulated condition and the subtraction of normal from emulated condition for the first participant in the left columns and for the second participant in the right columns.

Lower DTW values were found for both participants on the third, fourth, and fifth formants when compared to the first and second, indicating a high similarity between normal and emulated conditions. Hence, one can assume only a shift in the distribution of these formants. We found the amount of shift from the peaks in cross-correlation and shifted the emulated distribution and subtracted the regular from shifted emulated [Figure 4-15]. Now if we shift back the subtracted values, baselining regular condition, positive values show us the new frequencies present in the emulated and negative values show the missing frequencies in the emulated as are shown in [Figure 4-16].

4.3 MFCC

Mel Frequency Cepstral Coefficient (MFCC) is a widely used feature extraction technique in the field of speech and audio processing. Both have application for speech recognition, speaker identification, and audio classification. Several steps are required to compute MFCCs. First, audio signals must be segmented into short frames (e.g., 20 to 50 ms) to account for temporal variations in speech and Hamming widow is applied to them. Next, Fourier transform is applied to each frame to convert it from time domain to frequency domain, thus capturing the signal's spectral information. The Mel scale filter bank is then applied to mimic the human ear's non-linear perception of pitch, which emphasizes the importance of low over high frequencies. The next step is to compute the logarithm of the energy in each Mel frequency band, acknowledging the human ear's logarithmic (or log) perception of amplitude. Finally, Discrete Cosine Transform (DCT) is applied to the log energies, resulting in a set of uncorrelated coefficients that compactly represent the audio signal's spectral shape.

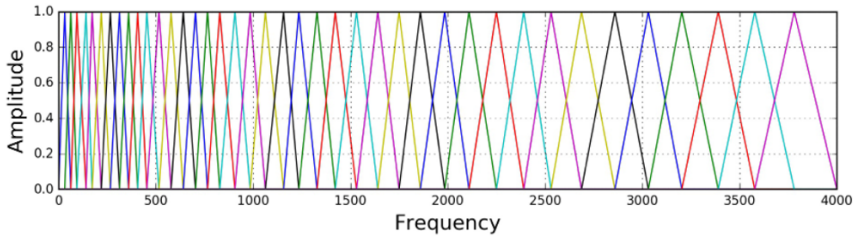

Figure 4-17 MFCC Filter Banks [12].

MFCC's coefficients are uncorrelated due to the definition of discrete cosine transform, which makes the coefficients orthogonal to each other:

$$C_n = \sum_{k=0}^{K-1} S(k) * cos\left[\frac{\pi n}{K}\left(k + \frac{1}{2}\right)\right] \qquad 4\text{-}17$$

in which, C(n) is the n^{th} cepstral coefficient; S(k) is the log Mel spectrum at the k^{th} band; and K is the total number of Mel frequency bands. To prove that coefficients are uncorrelated, their inner product must be calculated:

$$\sum_{k=0}^{K-1} cos\left[\frac{\pi n}{K}\left(k + \frac{1}{2}\right)\right] * cos\left[\frac{\pi m}{K}\left(k + \frac{1}{2}\right)\right] = 0 \quad for \quad m \neq n \qquad 4\text{-}18$$

MFCCs was applied to the audio signal of specific words that were repeated in "the caterpillar" passage. For improved visualization in a 3-dimensional space, MFCCs were further processed by applying PCA. Results demonstrate good separation between normal and emulated data for both participants. Visualized separation was verified with ANOVA analysis in which the low p-value rejects the null hypothesis.

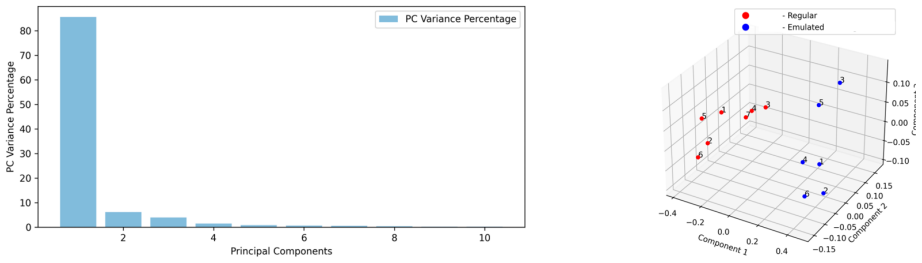

Figure 4-18 First participant's PCA visualization of the normal and emulated with obstacle for the word "caterpillar" pronunciations with the F-value of 11.77 and P-value of 0.0008 from ANOVA.

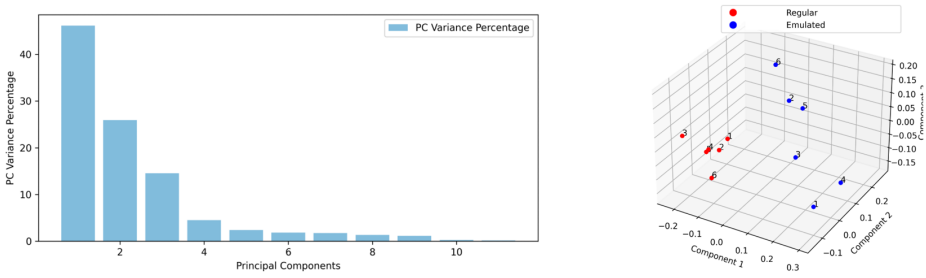

Figure 4-19 Second participant's PCA visualization of the normal and emulated with obstacle for the word "caterpillar" pronunciations with the F-value of 9.87 and P-value of 0.002 from ANOVA.

4.4 Summary

This chapter delved into the analysis of speech audio signals and the importance of methodology for extracting/modeling various audio features. We highlighted audio recognition software used to segment audio files into words and to detecting their time intervals. This process is critical for analyzing speech characteristics on a per-word basis. Various audio features (e.g., time duration, RMS sum, zero crossings, pitch or [fundamental frequency], and formant frequencies) were defined and discussed, including extraction methods for each. F0 distribution utilizing signal processing methods were detailed to explain how to compare normal and restrictive conditions, which aids in characterizing frequency loss, gain, and shifts. This information is important for understanding the impact of conditions like oral surgery on speech. Preliminary results demonstrated a major shift of frequencies for the third, fourth, and fifth formant for both study participants.

5 Machine Learning

This chapter discusses the methods and preliminary results for tracking patients' improvements during speech therapy following oral cancer surgery. First, the use of principal component analysis (PCA) for projecting feature space results is detailed, along with information about separating participant recordings and comparing normal and emulated conditions. Next, distances between normal and emulated data are detailed. Finally, machine learning models are introduced as a method for a binary classification for normal and emulated conditions. These are beneficial for monitoring speech rehabilitation after cancer treatment.

5.1 Principle Component Analysis

PCA is well-known method for reducing dimensionality and visualizing data. This statistical technique facilitates pattern identification by projecting original features into a new set of orthogonal vectors (or principal components), which are ordered by the eigen values' magnitude of covariance matrix from highest to lowest. This method was used to investigate the possibility of visualizing feature spaces in a lower dimension and finding separation between them.

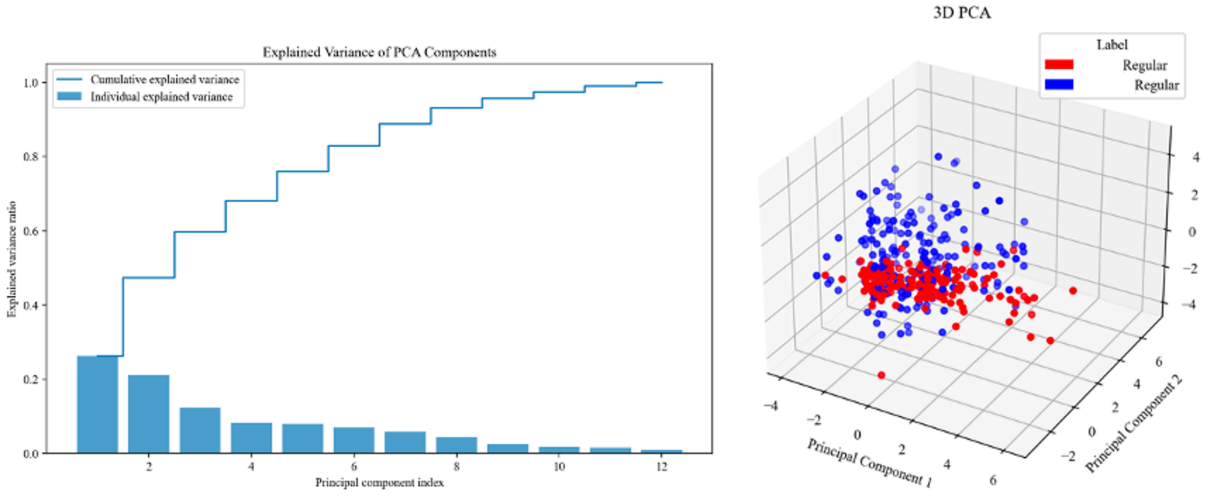

Figure 5-1 PCA Results of Feature Spaces for First Subject (Red) vs Second Subject (Blue), both under Normal Condition.

PCA 3D visualization and recognizing that these three components contain only 60 percent of variance implies that a good separation is not possible for two participants based on just three components. The low variance ratio from Principal Components (PCs) suggests some key points about the dataset and the nature of the variance it contains. Regarding its multidimensional nature, the feature space does not contain a small subset of features for capturing most variance. Instead, variance is spread out across many dimensions, implying that the dataset is complex and the information it contains cannot be easily summarized by just a few principal components. Regarding feature importance, PCA's explained variance ratio of a component indicates how much of the total variance in the dataset is captured by a specific component. Low percentages mean that each component captures only a small part of the total variance, suggesting that more than a small subset of features dominates the dataset (i.e. all the features contribute to the overall variance). In other words, low and gradually decreasing variance ratios indicate that the dataset structure used for this study is complex with many features contributing to its variance. This complexity necessitates using a larger number of principal components for a fuller representation of the data and implies that insights will likely come from understanding how groups of features interact, rather than focusing on just a few key features.

It is important to note that PCA loadings are coefficients of the linear combination that defines each principal component in terms of its original features. A high absolute value of a loading for a particular feature indicates that the feature has a strong influence on the principal component. A positive loading indicates that as the feature value increases, the principal component value tends to increase. Conversely, a negative loading indicates that as the feature value increases, the principal component value tends to decrease. For the first PC, the four formants, namely F_2 to F_5, have the highest contributions, likely due to the overall distinguishable patterns characterizing

the way that the two participants a) pronounce the context of “the caterpillar” passage, b) shape their vocal tract, and c) let the air flow to produce these frequencies. The second PC is mostly influenced by duration, RMS sum (or loudness), lip maximum vertical distance, and lip minimum horizontal distance. On its own, the RMS value is only related to audio signal volume; however, the definition for RMS sum makes RMS correlated with the duration of a word pronunciation, since it reflects summing the RMS value for small time frames of 20 ms to equal the total duration of a word with the correlation demonstrated in the second PC loadings since both have a similar effect on PC 2. Also, features related to lip motions are distinguishable patterns. Recall the example in the facial measurement chapter. The minimum horizontal distance of the lips (lip width) was different for the two participants. One showcased a shrinking effect; the was nearly solid. The high absolute value of magnitude for min distance horizontal indicates this differentiating pattern. Other PC loadings can be interpreted using the same approach. The following figures highlight PCA comparisons between normal and emulated conditions. The results for both participants are similar to comparisons for low variance ratio and PC loading coefficients.

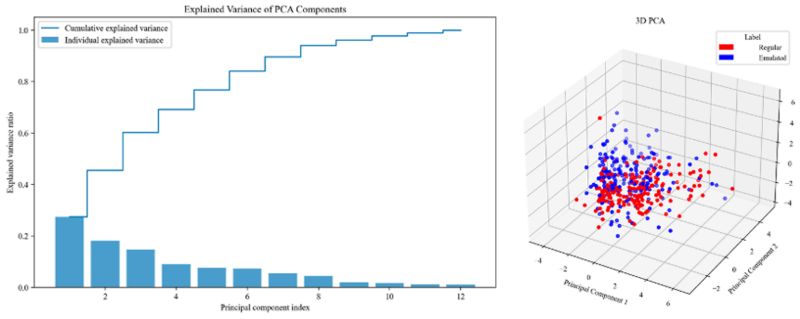

Figure 5-2 First participant under normal condition vs. first participant under emulated condition with oral obstacle.

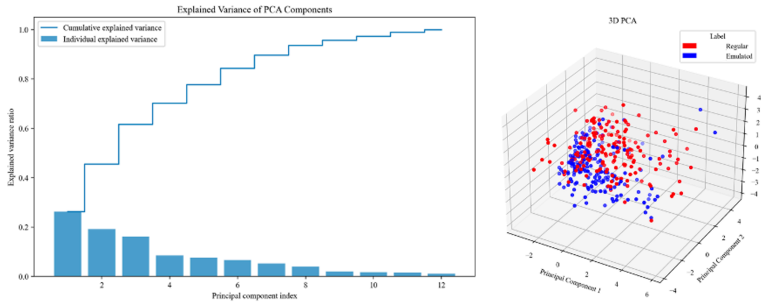

Figure 5-3 Second participant under normal condition vs. second participant under emulated condition with oral obstacle.

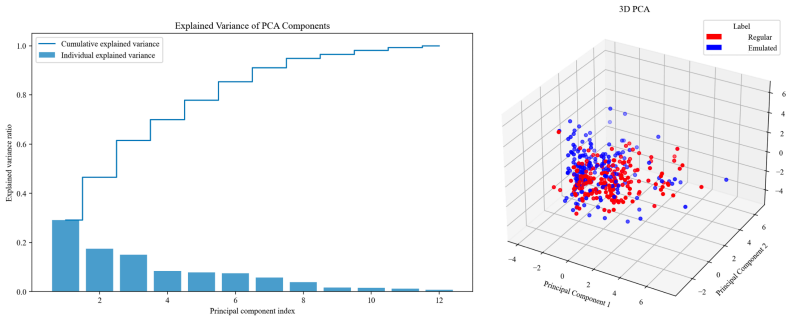

Figure 5-4 First participant under normal condition vs. first participant under emulated condition with clenched teeth.

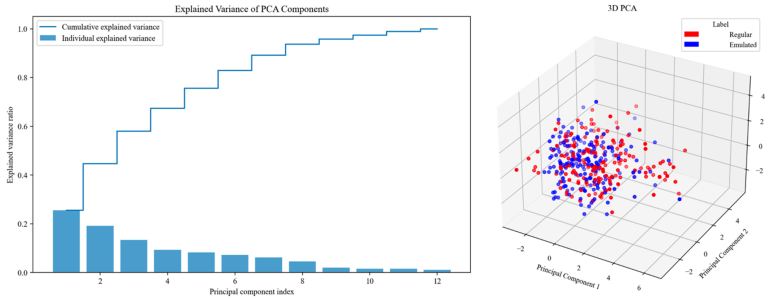

Figure 5-5 Second participant under normal condition vs. second participant under emulated condition with clenched teeth.

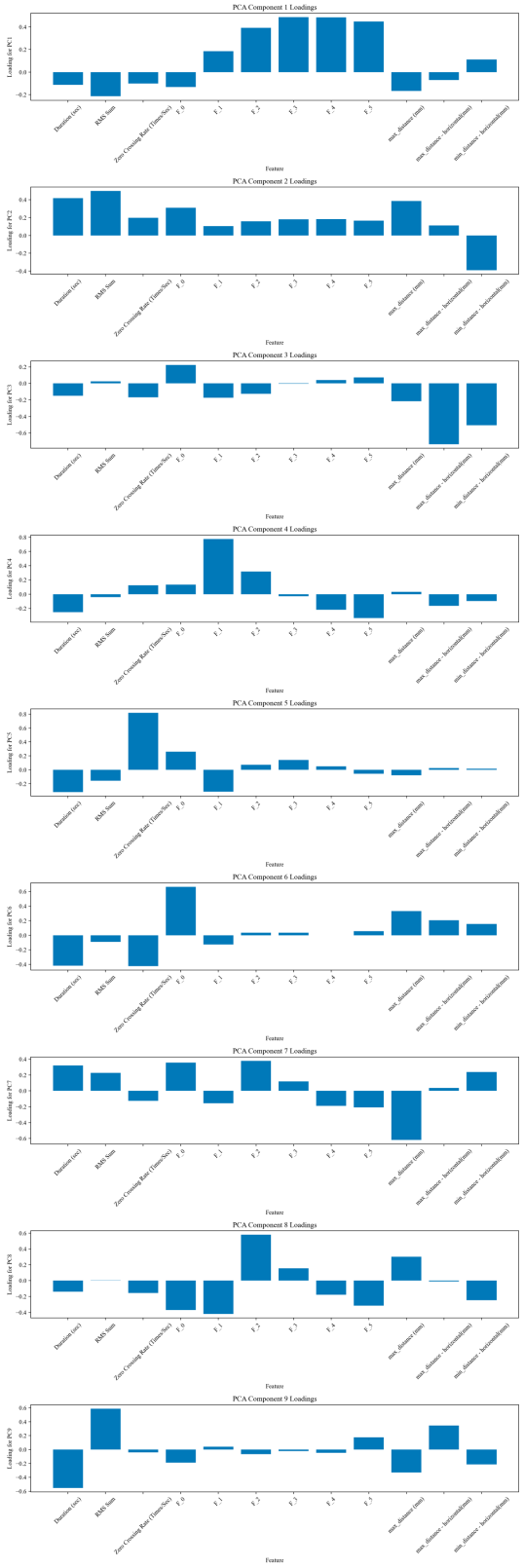

Figure 5-6 PCA component loadings, first participant vs. second participant, both under normal condition.

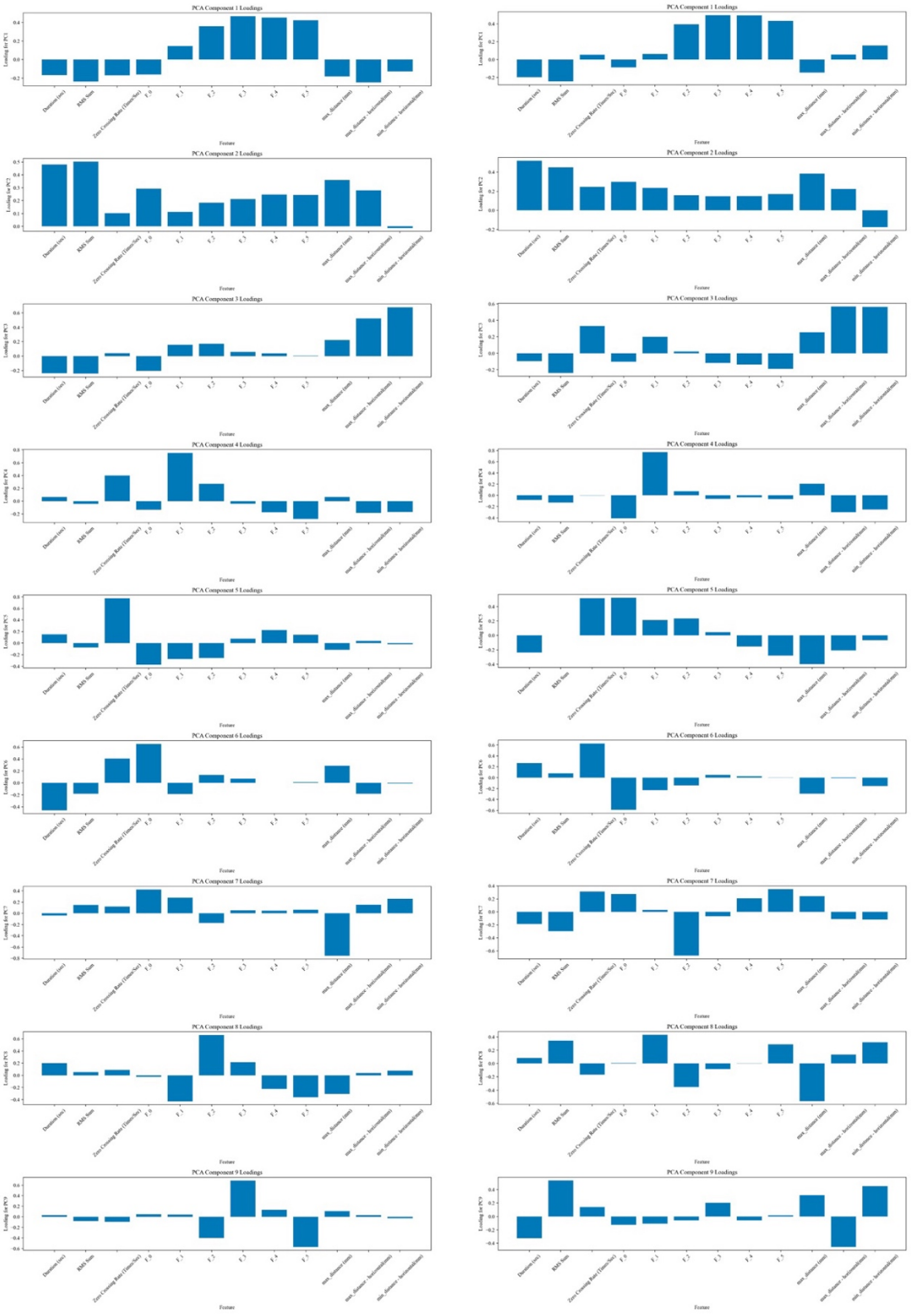

Figure 5-7 PCA component loadings., left: First participant under normal condition vs. first participant under emulated condition with oral obstacle under tongue. Right: Second participant under normal condition vs. second participant under emulated condition with oral obstacle under tongue.

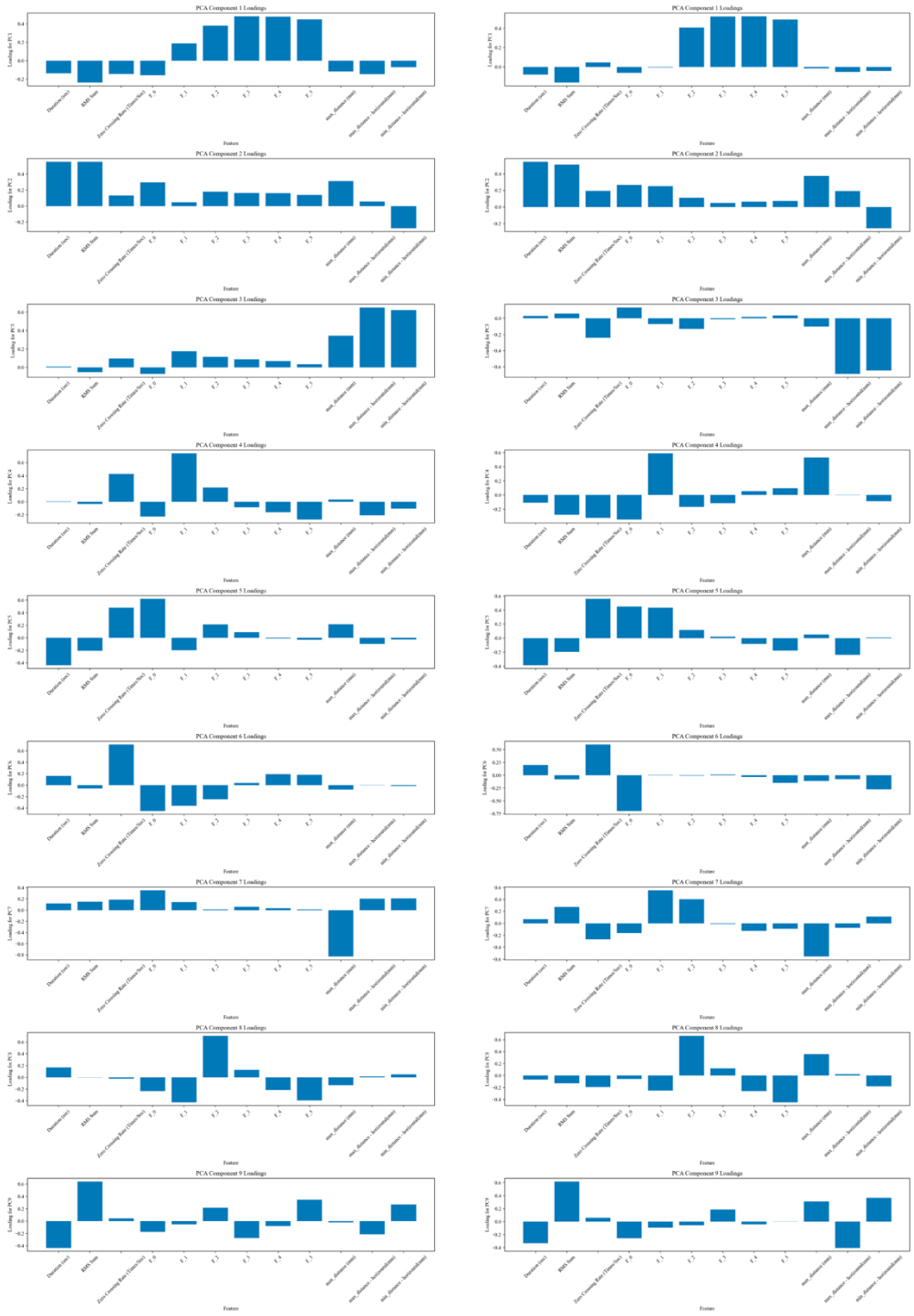

Figure 5-8 PCA component loadings. Left: First participant under normal condition vs. first participant under emulated condition with teeth clenched together. Right: Second participant under normal condition vs. second participant under emulated condition with teeth clenched together.

5.2 Feature Space Distance Analysis

In speech therapy it is important to have a measurable method for tracking patient progress, particularly following surgical interventions that impact vocalization. The method proposed in this work is a data-driven metric based on feature space (i.e., the distance between normal (pre-surgery) and emulated (post-surgery) conditions. Speech data gathered before surgery provides a patient-specific ground truth, reflecting the individual's inherent linguistic traits and motor capabilities. This personalized baseline is important, as it considers natural variability in speech, and then guides targeted interventions for post-operative recovery. Quantitative analysis provides a suitable qualitative assessment, as it facilitates a precise evaluation of speech characteristics. These include a spectrum of features, like formant frequencies, temporal aspects, and articulatory motion dynamics, all of which are integral to the components of speech. By assessing the multidimensional nature of these attributes, the complexity of speech production and its deviations from normal pre-surgery speech can be captured. Normalization of feature spaces is key for meaningful comparisons that integrate all features and ensure that each contributes to the distance in the same way. For normalization, feature vectors are rescaled when each is divided by the maximum sum of corresponding features across the recordings—both normal and emulated—given that the Manhattan distance was calculated between the two datasets.

This method allows therapists to set measurable and individualized therapeutic goals, thereby customizing interventions that address specific deficits identified via feature space analysis. Additionally, the evolution of these distances over time provides an objective metric for tracking improvements and progress toward pre-surgery articulation.

Table 5-1 Feature space distances

Comparison Cases	Manhattan Distance
First Subject's Normal Condition and Emulated with Oral Obstacle	3.49
First Subject's Normal Condition and Emulated with Clenched Teeth	3.71
Second Subject's Normal Condition and Emulated with Oral Obstacle	4.12
Second Subject's Normal Condition and Emulated with Clenched Teeth	3.18

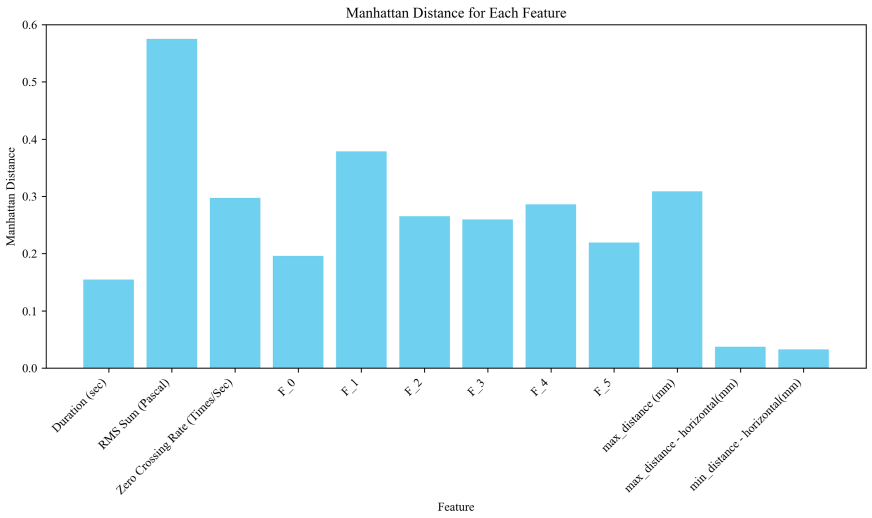

Figure 5-9 First participant's normal condition and emulated with oral obstacle.

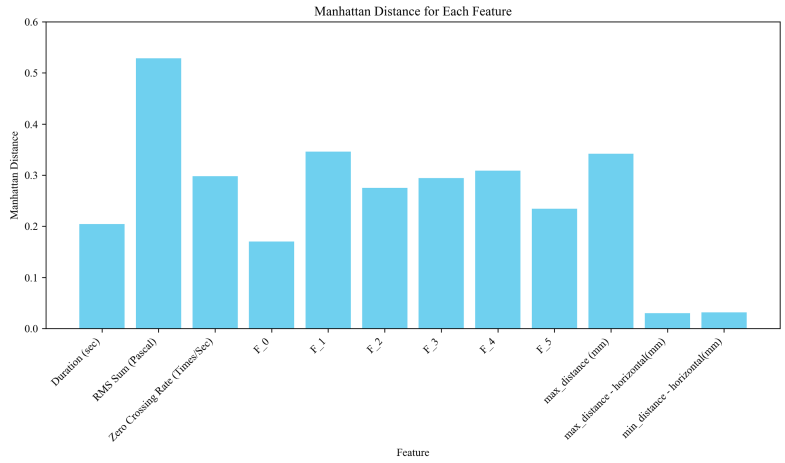

Figure 5-10 First participant's normal condition and emulated with clenched teeth.

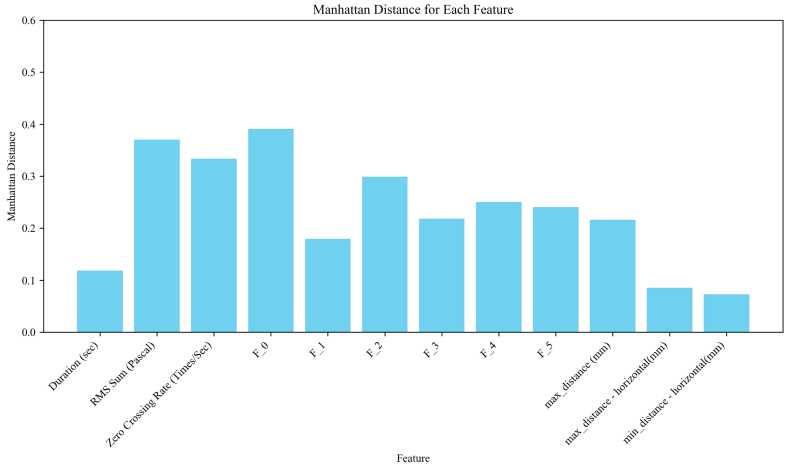

Figure 5-11 Second participant's normal condition and emulated with oral obstacle.

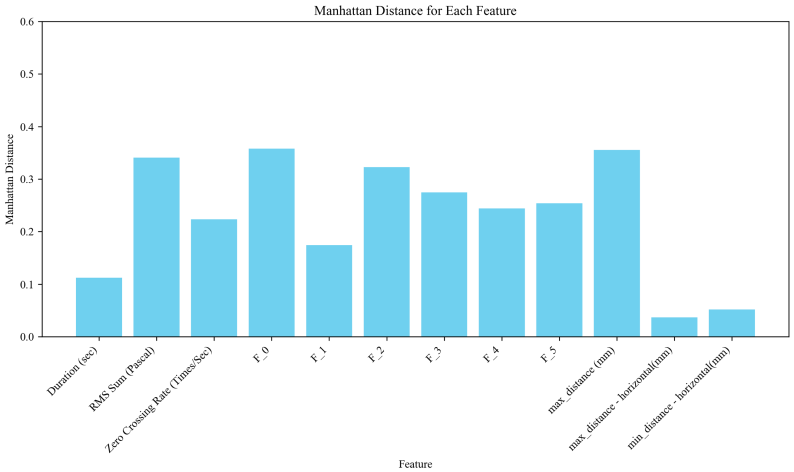

Figure 5-12 Second participant's normal condition and emulated with clenched teeth.

Distances highlighted in Figures 5-9, 5-10, 5-11, and 5-12 demonstrate a low value for both the minimum and maximum horizontal distance (i.e., lip's width) This finding was intuitively predicted, as emulations did not limit the sides of lips and predicted similar. The maximum distance of lips changed for both participants under the restrictive conditions (see distributions in Figure 3-7.) and resulted in either a higher contribution to maximum vertical distance (i.e., max_distacne) or a maximum distance between the upper and lower lips. The ZCR and five formants—F_1 to F_5—are characterized with a fairly high distance due to the effect of physical restrictions on articulation. There was not a big difference for time duration when compared to a normal condition, which resulted in the lower contribution to distances for this feature.

5.3 Machine Learning Classification

This section explores distinctions between normal and emulated speech conditions, as well as variances across different participants, Machine learning classification models were used to understand the impact of oral constraints on speech production. This approach proved crucial for identifying specific speech characteristics influenced by emulated post-operative conditions, which offered insights for speech therapy and rehabilitation techniques. This process can demonstrate the stage of articulation improvement and assist speech therapists at each step of the treatment. Moreover, analyzing speech variations among individuals enables the customization of therapeutic interventions, ensuring they are applicable to personal speech patterns and recovery needs.

The developed model utilized for this classification is termed random forest with 100 trees, Maximum number of leaf nodes is 10, and number of features randomly sampled for each split is set with the default value of a scikit-learn library (i.e., equal to the square root of the number of input features or 12. Each row in the feature space was labeled either normal or emulated in the

corresponding datasets. During testing, one emulated scenario was concatenated with a patients' normal condition. 80 % of collected data was used for the training set and 20 percent for the test set to predict a binary classification. Given the small dataset, data had to be split randomly 100 times, and trained to establish a new model. Reported results are the average of Figure 5-9 and Figure 5-10.

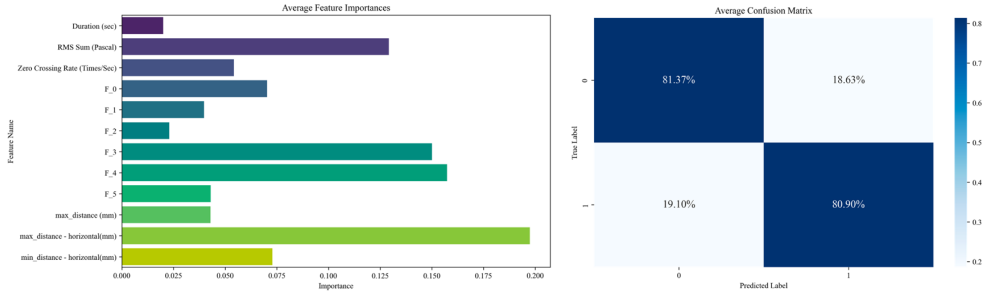

Figure 5-13 Classification between normal condition vs. oral obstacle under the tongue.

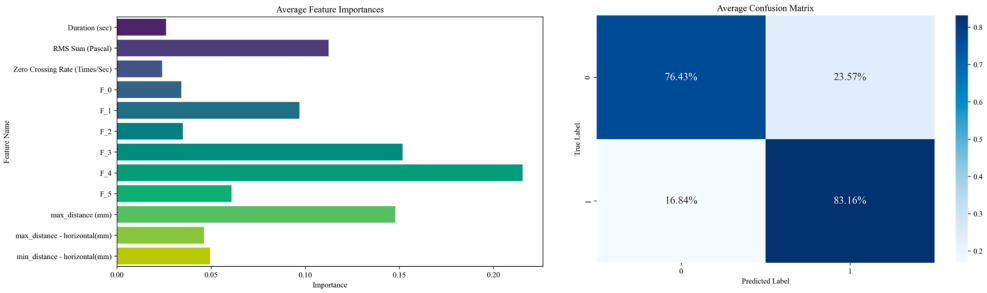

Figure 5-14 Classification between normal condition vs. teeth clenched together.

5.3.1 Feature Importance

Feature importance was extracted utilizing scikit-learn library attributes. In both classification models, there was a high importance for separating normal conditions from emulation data in either the third or fourth formants, which demonstrates that either having an oral obstacle under the tongue or clenching teeth together while speaking has a significant articulation effect on the two formants. It is worth reiterating that in the narrative in Chapter 4 highlighted a clear shift in the

fourth formant when analyzing windowed formant frequencies (see Figure 4-11). The “RMS sum” (i.e., energy fo audio signl) proved to be the next audio related feature causing acceptable separation across different cases The reason for this phenomenon is that when vocal tract motion is limited by either an oral obstacle under the tongue or clenched teeth, airflow through the vocal tube is changed, which in turn changes the energy or volume of speech.

The lip motion-related feature relative to maximum horizontal distance and a side shrinking effect of the lips, proved to create high distinction between normal and emulated condition, given the presence of an oral obstacle [(see Figure 5-9]). Chapter 3 results demonstrated that when analyzing the distribution of this same feature there was a difference in mean and standard deviation. This also demonstrated an effect in the classification model. Results for classification when evaluating the difference between a normal condition and speaking while teeth were clenched proved that maximum lip distance is highly important for classification. For example, when teeth are clenched, distance between the lips is limited, since the upper and lower jaws are fixed, which prevents the mouth from opening freely.

5.4 Summary

This chapter delved into the methodologies and preliminary findings for monitoring patient progress during speech therapy following cancer surgery. A feature space projection analysis utilizing Principal Component Analysis (PCA) assessed the distinction between participants' speech recordings under both normal conditions and post-surgery simulations in a low dimensional space. Results suggest that PCA addresses dataset complexity and indicates that rather than relying on a small subset of features, it is best to spread variance across many dimensions. This complexity signifies that a more substantial number of principal components is necessary for full data representation. Further, the importance of PCA loadings was discussed, which highlighted the

influence of specific features on principal components and indicated significant patterns in speech production differ among participants. Contributions of formant frequencies and articulatory dynamics to these patterns are particularly notable. Latter sections focused on distance analysis in feature spaces, proposing a data-driven metric to evaluate speech therapy progress. The measure compared the distances between pre- and post-surgery speech data and offered a personalized baseline for monitoring recovery and establishing measurable therapeutic goals. Normalization of feature spaces ensured equitable comparisons across different speech characteristics. This chapter also outlined the application of machine learning models for binary classification of normal and emulated speech conditions among participants. This approach identifies speech production characteristics affected by oral constraints and encourages superior speech therapy and rehabilitation strategies. Models highlighted the significance of certain speech features, including formant frequencies and articulatory dynamics, when distinguishing between normal and simulated post-operative conditions. In summary, this chapter provided a comprehensive analysis of speech feature space for monitoring patient recovery in speech therapy following oral cancer surgery. It also promoted the use of PCA for understanding data complexity, executing distance analysis for progress tracking, and leveraging machine learning for detailed speech characteristic classification.

6 Conclusion and Future Work

The research presented in this thesis details a variety of methods for characterizing audio and facial motion during speech tasks. The first step was analyzing the distribution of extracted F0s of small-time frames to discover behavioral patterns under both normal and emulated conditions. More specifically, normal condition were baselined, and then a dynamic time warping algorithm was used to find the minimal distance between emulated formants and typical ones to establish an awareness of the similarity between them. Results from two participants reflected a high similarity for two conditions on the fourth and fifth formants. Cross-correlation indicated the amount of formants shift. Coupling this informative data with the characterization of frequency shift indicated newly generated and lost frequencies, which was explained in Chapter 4. Another approach leveraged the window size of the proposed formant derivation algorithm with the length of a spoken word to archive the information for a machine learning feature space. Utilizing these frequencies and other audio and motion related features, this work's analysis demonstrated a distance between a participant's baseline and emulated conditions, which can be used to track improvement during motor rehabilitation and speech therapy after cancer treatments (e.g. surgery, chemotherapy, and the like). Moreover, developed machine learning models classified differences among participants and speech impairments, which makes these technologies a powerful source for tracking the stage of rehabilitation relative to that being trained on a large dataset. In conclusion, the research and experimental results detailed in this thesis lay the groundwork for further advancing analysis methodology and improve tracking the HNC patient speech motor task rehabilitation/progress following cancer treatment. This work could minimize the duration of a patient's medical care and improve their quality of life.

6.1 Future Directions

This chapter highlights various directions for future research. Of utmost interest is estimating vocal tract shape and tongue position, which could provide insightful information for medical doctors and speech therapists. The method developed in [28] generates vocal tract shape as observed in X-ray images of most English vowels by utilizing the first three formant frequencies. Using the methodology proposed herein, facial border tracking and component locations (e.g., lips), will add valuable information. This advancement will lead to the development of a more advanced algorithm to integrate motion and speech audio formants, and then more accurately estimate vocal tract shape and its dynamics.

Another idea that could further expand the research carried out for this thesis is adding a lateral camera and developing an algorithm to track a selected points on one side of a patient's head. A particular methodology could then be developed to synchronize front and lateral camera recordings. Adding insertion dynamics of head lateral motion to the vocal tract shape estimation algorithm could provide valuable information for reconstructing a 3D shape of the vocal tract. The process should involve depth estimation techniques and triangulation methods to merge 2D data from each camera into a cohesive 3D structure. Finally, the developed methodologies must be validated through clinical trials and/or studies involving participants with and without speech impairments. Such authentication is poised to refine algorithms and confirm the utility of 3D models for diagnosing, understanding, and treating speech generation issues. While existing methods are able to estimate vocal tract shape with high accuracy, the work in [29] utilized high frame rate (166/sec) MRI scanning of a participant's head during the speech task with a spatial resolution of 2.2 * 2.2 * 5.0 mm^3. However, limited accessibility to MRI equipment makes this method impractical for widespread research or clinical use. Using the proposed approach with

synchronized video recordings from frontal and lateral cameras presents a more accessible and cost-effective solution.

7 References

[1] Clarke P, Radford K, Coffey M, Stewart M. Speech and swallow rehabilitation in head and neck cancer: United Kingdom National Multidisciplinary Guidelines. The Journal of Laryngology & Otology. 2016 May;130(S2):S176-80.

[2] Murphy BA, Gilbert J. Dysphagia in head and neck cancer patients treated with radiation: assessment, sequelae, and rehabilitation. In Seminars in radiation oncology 2009 Jan 1 (Vol. 19, No. 1, pp. 35-42). WB Saunders.

[3] Weber C, Dommerich S, Pau HW, Kramp B. Limited mouth opening after primary therapy of head and neck cancer. Oral and maxillofacial surgery. 2010 Sep;14:169-73.

[4] List MA, Bilir SP. Functional outcomes in head and neck cancer. In Seminars in radiation oncology 2004 Apr 1 (Vol. 14, No. 2, pp. 178-189). WB Saunders.

[5] van der Molen L, van Rossum MA, Jacobi I, van Son RJ, Smeele LE, Rasch CR, Hilgers FJ. Pre-and posttreatment voice and speech outcomes in patients with advanced head and neck cancer treated with chemoradiotherapy: expert listeners' and patient's perception. Journal of Voice. 2012 Sep 1;26(5):664-e25.

[6] Starmer H, Sanguineti G, Marur S, Gourin CG. Multidisciplinary head and neck cancer clinic and adherence with speech pathology. The Laryngoscope. 2011 Oct;121(10):2131-5.

[7] Jin H, Liao S, Shao L. Pixel-in-pixel net: Towards efficient facial landmark detection in the wild. International Journal of Computer Vision. 2021 Dec;129(12):3174-94.

[8] Bain M, Huh J, Han T, Zisserman A. Whisperx: Time-accurate speech transcription of long-form audio. arXiv preprint arXiv:2303.00747. 2023 Mar 1.

[9] Tom Bäckström, Okko Räsänen, Abraham Zewoudie, Pablo Pérez Zarazaga, Liisa Koivusalo, Sneha Das, Esteban Gómez Mellado, Mariem Bouafif Mansali, Daniel

Ramos, Sudarsana Kadiri and Paavo Alku “Introduction to Speech Processing”, 2nd Edition, 2022. URL: https://speechprocessingbook.aalto.fi, DOI: 10.5281/zenodo.6821775.

[10] Rabiner LR, Juang BH. Fundamentals of speech recognition. Tsinghua University Press; 1999.

[11] Eliathamby Ambikairajah “Introduction to Speech Processing” http://eemedia.ee.unsw.edu.au/contents/elec9344/LectureNotes/

[12] Speech Processing for Machine Learning https://haythamfayek.com/2016/04/21/speech-processing-for-machine-learning.html

[13] Zhang S, Zhu X, Lei Z, Shi H, Wang X, Li SZ. Faceboxes: A CPU real-time face detector with high accuracy. In2017 IEEE International Joint Conference on Biometrics (IJCB) 2017 Oct 1 (pp. 1-9). IEEE.

[14] FaceBoxesV2, https://github.com/jhb86253817/FaceBoxesV2?tab=readme-ov-file

[15] Olson ML, Shedd DP. Disability and rehabilitation in head and neck cancer patients after treatment. Head & neck surgery. 1978 Sep;1(1):52-8.

[16] Samuel SR, Maiya AG, Fernandes DJ, Guddattu V, Saxena PP, Kurian JR, Lin PJ, Mustian KM. Effectiveness of exercise-based rehabilitation on functional capacity and quality of life in head and neck cancer patients receiving chemo-radiotherapy. Supportive Care in Cancer. 2019 Oct 1;27:3913-20.

[17] Radford, A. et al. Robust speech recognition via large-scale weak supervision. Preprint at https://arxiv.org/abs/2212.04356 (2022).

[18] whisperX, https://github.com/m-bain/whisperX

[19] PIPNet, https://github.com/jhb86253817/PIPNet

[20] https://www.nidcr.nih.gov/health-info/oral-cancer

[21] Huang SH. Oral cancer: Current role of radiotherapy and chemotherapy. Medicina oral, patologia oral y cirugia bucal. 2013 Mar;18(2):e233.

[22] Klussmann JP, Schädlich PK, Chen X, Rémy V. Annual cost of hospitalization, inpatient rehabilitation, and sick leave for head and neck cancers in Germany. Clinic Economics and Outcomes Research. 2013 May 16:203-13.

[23] Blomberg M, Nielsen A, Munk C, Kjaer SK. Trends in head and neck cancer incidence in Denmark, 1978–2007: focus on human papillomavirus associated sites. International journal of cancer. 2011 Aug 1;129(3):733-41.

[24] Robert Koch Institute, Society of Cancer Registries in Germany. Krebs in Deutschland 2007/2008 [Cancer in Germany 2007/2008]. Berlin: Robert Koch Institute, Society of Cancer Registries in Germany; 2012. German.

[25] Guntinas-Lichius O, Wendt T, Buentzel J, et al. Head and neck cancer in Germany: a site-specific analysis of survival of the Thuringian cancer registration database. J Cancer Res Clin Oncol. 2010;136(1): 55–63.

[26] Lee JM, Turini M, Botteman MF, Stephens JM, Pashos CL. Economic burden of head and neck cancer. A literature review. Eur J Health Econ. 2004;5(1):70–80.

[27] St Guily JL, Borget I, Vainchtock A, Rémy V, Takizawa C. Head and neck cancers in France: an analysis of the hospital medical information system (PMSI) database. Head Neck Oncol. 2010;2:22.

[28] Ladefoged P, Harshman R, Goldstein L, Rice L. Generating vocal tract shapes from formant frequencies. The Journal of the Acoustical Society of America. 1978 Oct 1;64(4):1027-35.

[29] Fu M, Barlaz MS, Holtrop JL, Perry JL, Kuehn DP, Shosted RK, Liang ZP, Sutton BP. High-frame-rate full-vocal-tract 3D dynamic speech imaging. Magnetic resonance in medicine. 2017 Apr;77(4):1619-29.

[30] Patel R, Connaghan K, Franco D, Edsall E, Forgit D, Olsen L, Ramage L, Tyler E, Russell S. “The Caterpillar”: A novel reading passage for assessment of motor speech disorders.