SPEAKER-SPECIFIC ADAPTATION OF MAEDA SYNTHESIS PARAMETERS FOR AUDITORY FEEDBACK

by

Joseph Vonderhaar, B.S.

A Thesis submitted to the Faculty of the Graduate School,
Marquette University,
in Partial Fulfillment of the Requirements for
the Degree of Master of Science

Milwaukee, Wisconsin

May 2017

ABSTRACT
SPEAKER-SPECIFIC ADAPTATION OF MAEDA SYNTHESIS PARAMETERS FOR AUDITORY FEEDBACK

Joseph Vonderhaar, B.S.

Marquette University, 2017

The Real-time Articulatory Speech Synthesizer (RASS) is a research tool in the Marquette Speech and Swallowing lab that simultaneously collects acoustic and articulatory data from human participants. The system is used to study acoustic to articulatory inversion, articulatory to acoustic synthesis mapping, and the effects of real-time acoustic feedback. Electromagnetic Articulography (EMA) is utilized to collect position data via sensors placed in a subject's mouth. These kinematic data are then converted into a set of synthesis parameters that controls an articulatory speech synthesizer, which in turn generates an acoustic waveform matching the associated kinematics. Independently from RASS, the synthesized acoustic waveform can be further modified before it is returned to the subject, creating the opportunity for involuntary learning through controlled acoustic feedback.

In order to maximize the impact of involuntary learning, the characteristics of the synthetically generated speech need to closely match those of the participant. There are a number of synthesis parameters that cannot be directly controlled by the subject's articulatory movements such as fundamental frequency and parameters corresponding to physiological measures such as vocal tract length and overall vocal tract size. The goal of this work is to develop a mechanism for automatically determining RASS internal synthesis parameters that provide the closest synthesis parameter match to the subject's acoustic characteristics, ultimately increasing the system's effect positive effect on involuntary learning.

The methods detailed in this thesis examine the effects of altering both time-independent and time-dependent synthesis parameters to increase the acoustic similarity between the synthesized and real subject speech. The fundamental frequency and first two formant values are studied in particular across multiple vowels to determine the time-independent parameter settings. Time-dependent parameter analysis is performed through the use of a real-time parameter-tracking configuration. Results of this work provide a way of adapting the Maeda synthesis parameters in RASS to be speaker-specific and individualize the study of auditory feedback. This research will allow researchers to better customize the RASS system for individual subjects and alter involuntary learning.

ACKNOWLEDGEMENTS

Joseph Vonderhaar, B.S.

There are so many friends and professors I would like to thank for their contributions to my Marquette education, research, and development as an engineer that it would be difficult to name them all. However, there are a few that stand out in particular.

I would like to express my appreciation for my research advisor, Dr. Michael Johnson, who always has my best interests in mind. His supervision, knowledge, and insight made this thesis possible. I wish him the continued success in his new role at the University of Kentucky.

I would also like to thank Dr. Jeffrey Berry and the researchers in Marquette's Speech and Swallowing Lab. Their flexibility and guidance was instrumental to my thesis.

I would like to extend a thank you to the members of my thesis committee for the time they spent reviewing my work and for their valuable feedback.

Finally, I would like to express my sincere gratitude to my family who has kept me motivated through this process. In particular, I would like to thank my sister, Kelly, who through the past years has been an invaluable friend, classmate, and fellow researcher. This thesis would not have been possible without their support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1 Opening

Speech disorders affect a significant number of people in the United States. Somewhere between six and eight million people suffer from a speech impairment [1]. Dysarthria, which is one of these disorders, is a result of damaged neural mechanisms which are used to control speech. More specifically, damaged mechanisms can cause changes in articulatory movements which often lead to mispronunciations and deviated speech acoustics. Articulatory impairment often comes in the form of movement reduction, slowness, and poor coordination [2]. A current problem associated with this disorder is the lack of effective rehabilitative therapies for people trying to recover and improve their pronunciation. One related area of research is involuntary acoustic learning, where modified kinematic-driven acoustic feedback is used to alter subjects' articulatory movements. Marquette University's Speech and Swallowing Lab has conducted several studies investigating such involuntary sensorimotor learning [3].

Marquette uses an Electromagnetic Articulography (EMA) system to acquire kinematic data from subjects. These data are then fed into a software system for speech synthesis and ultimately acoustic feedback. The software system, Real-time Articulatory Speech Synthesizer (RASS), maps kinematic data from sensors to acoustic synthesis parameters [4]. These synthesis parameters represent both pronunciation related components such as tongue shape and movement as well as physiological components such as vocal tract length and fundamental frequency (F0). Due to physical subject variability, the synthesis parameters related to physiological components necessarily vary

substantially from person to person. These synthesis parameters are not currently controlled by the RASS system, but are essential to enable RASS to match individual acoustic characteristics. The goal of this research is to develop and test methods to best match the RASS synthesizer to individual subjects.

1.2 Speech Production

The focus of this thesis is centered on synthesizing speech and subject matching, so is it important for one to understand how sound is defined and developed into speech. Sound is essentially a pressure wave created from the compression and rarefaction of surrounding air molecules. The longitudinal wave is parallel to the energy applied and can be modeled by a sine wave. The peaks of the wave represent maximum compression, and the troughs represent moments of maximum rarefaction. Speech, one form of sound production, is generated by air-pressure waves oscillating through the mouth and nostrils of a human. Within speech, phonemes are considered the most basic units and can be grouped into two categories, consonants and vowels. The difference between these two groups is the presence of constrictions or obstructions in the throat during articulation. Vowels are articulated without significant impediments, while consonants rely on constrictions or obstructions during speech [5].

The human speech apparatus consists of several key components. The source of the system is the lungs, from which air is forced through the trachea, across the vocal folds, and to the larynx. The vocal folds stretch across the larynx from back to front and join at the glottis, controlling the air flow from the lungs. From the larynx, the velum, or soft palate, allows air to pass through the nasal cavity or mouth, acting like a valve. The

air that passes through the nasal cavity and the mouth is filtered by articulators which are used to regulate the sound and ultimately turn it into speech. Voiced sounds are the focus of this research and are produced by the vibrations that occur when air passes through closed vocal folds. The tension of the vocal folds and the resulting air pressure form a glottal excitation signal that then passes through the articulators. Unvoiced sounds occur when the vocal folds don't vibrate together [5].

The hard palate, which is the roof of the mouth, is used for articulation in conjunction with the tongue, which is a flexible articulator. Teeth are also important to speech production, specifically as a brace for the tongue to produce consonants. Lips, which are the final articulator before air exits the mouth, play a role in affecting vowel quality. They can be rounded for certain vowels or completely closed to stop the excitation of air [5].

During phonation, the rate of the cycling is called the fundamental frequency (F0) and is the main contributor to the perception of pitch. Although a vowel does not sound the same when generated at different fundamental frequencies, it often involves a similar envelope of harmonics [6]. The release of air from the lungs can be modeled as a glottal wave and analyzed as a sum of sine waves. When the vocal tract is simplified to a uniform tube with a uniform cross-sectional area, one end closed (at the glottis), and one end open (at the lips), any change in the shape of the vocal tract will change the resonances of the glottal wave too. The resonances that are typically the result of certain articulator alignments are called formants, concentrations of acoustic energy around a certain frequency. The first formant value, F1, is generally attributed to the open/closed

characteristics of the back of the mouth cavity. The second formant, F2, is related to the front/back position of the tongue [5].

The significance in studying formants in relation to this research is that no two people pronounce a vowel exactly the same. People have varying shapes and sizes of vocal tracts and articulators, which cause slightly different formants to be produced for the same vowels. By analyzing formant values for several vowels among diverse groups of people, one can achieve some form of individual identification which aids in the adaptation of speaker-specific synthesis parameters in RASS.

1.3 Research Objectives

The main objective of this research is to determine a more accurate way to match Maeda synthesis parameters to a subject's acoustic characteristics in RASS, essentially adapting the model to the subject. The vocal tract model used for speech synthesis in RASS consists of parameters that are fixed during synthesis (time-independent) and those that can vary in real-time (time-dependent). The objective of this thesis is to study how varying both types of parameters will produce the closest match between a subject's synthesized and real speech. More specifically, the time-independent parameters that control the laryngeal height and overall size of the vocal tract model will be studied in addition to the time-dependent fundamental frequency parameter. The combination of both approaches leads to a more accurate speaker-specific adaptation that can ultimately aid in the study of rehabilitative involuntary learning.

1.4 Overview of Thesis

The remaining portion of this thesis will be organized in to the following chapters: Background (Chapter 2), Time-Independent Parameter Matching (Chapter 3), Time-Dependent Parameter Matching (Chapter 4), and Conclusion (Chapter 5).

CHAPTER 2: BACKGROUND

2.1 System Overview

In order to match synthesis parameters to the acoustic characteristics of the subject, a better understanding of the RASS system is necessary. The main components of RASS include the NDI Wave system, mapping algorithms, and the Maeda synthesizer. The synthesizer is more commonly called VTDemo, which stands for Vocal Tract Acoustics Demonstrator. As Figure 1 shows, the acoustic signal is streamed into RASS from the human subject and the output is sent into Audapt, which is a tool used to alter the speech for specific learning outcomes. The signal is then fed back to the beginning of the system, where the subject can hear the speech through headphones and begin to correct pronunciation through fine-tuning motor behavior.

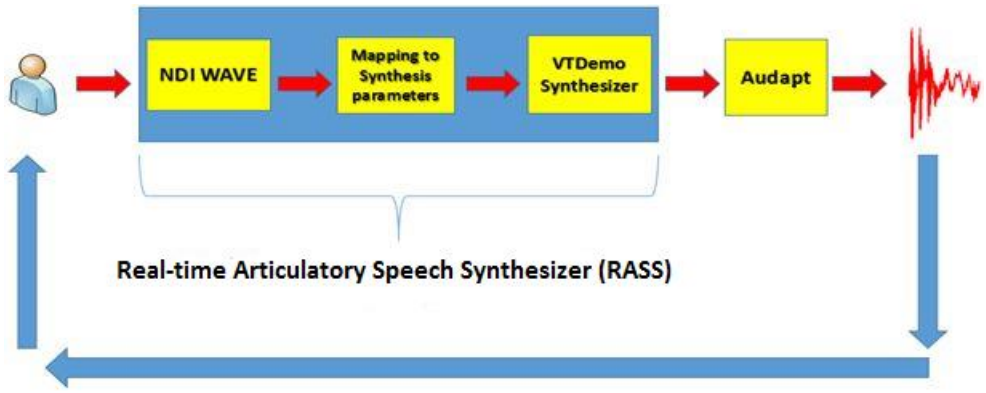

Figure 1: Lab Configuration Featuring RASS [5]

The purpose of the system in Figure 1 is to study involuntary learning through acoustic feedback. The vocal tract, as modeled in the VTDemo synthesizer, filters sound based on the positions of articulators. Therefore, the corresponding synthesis parameters need to be identified that best reflect those articulators and match the subject's acoustic

characteristics to the synthesized voice. In RASS, the speech synthesis parameters are determined by sensors that are placed on the subject's articulators. The kinematic data are gathered in real-time by the subject's sensors and then entered into the algorithm for mapping to synthesis parameters. After the appropriate synthesis parameters are generated which align with subject's acoustic characteristics, they are entered into the VTDemo speech synthesizer. Outside of this overview, there are several small calibrations and sensor alignments that occur outside of the simplified diagram in Figure 1. A breakdown from the software side of the system, the moment the data are collected until they leave RASS, can be seen in Figure 2.

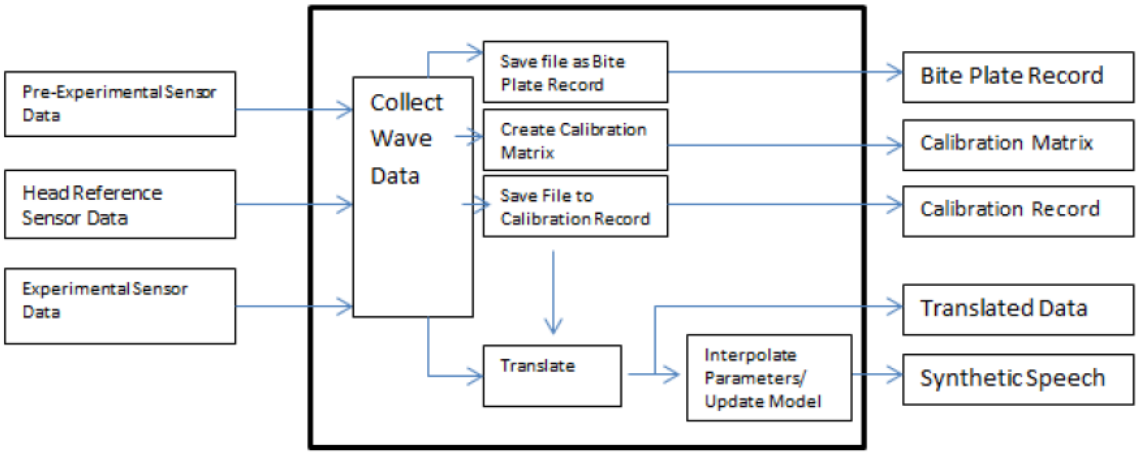

Figure 2: Software Breakdown of RASS [7]

2.2 NDI Wave System

2.2.1 About the NDI Wave System

The function of the NDI Wave system, which is to collect kinematic data from the human subject's articulators, is achieved through the use of electromagnetic articulography. The Wave System is described by NDI as "an electromagnetic non-line-

of-sight motion capture system" [7]. EMA works through the use of sensors that are attached to human articulators. A small, static electromagnetic field is then produced surrounding an individual's head to allow for sensor tracking in three dimensions. The signal in the sensors is induced through electromagnetic induction. As a subject speaks, the position and orientation of the sensors change and is reported to the data collection system, NDI Wave, in real-time.

Specific to the NDI Wave, the system consists of a box containing transmitter coils and a data collection component. Eight sensors are able to be tracked in two possible sizes of electromagnetic fields, either 300 mm^3 or 500 mm^3, which are offset from the front of the field generator by 40 mm. The accuracy of the system is within 0.5 mm, which is an acceptable tolerance for gathering kinematic data. The sampling rate for the standard system is 100 Hz but is able to be increased up to 400 Hz with an upgrade. Furthermore, the upgraded system, which is the unit Marquette's Speech and Swallowing Lab uses, can collect data from eight additional sensors. Figure 3 shows the NDI Wave System generator and the corresponding electromagnetic field that is generated during operation [8] [9].

The NDI Wave system, which tracks kinematics along the human vocal tract, consists of the following main components: a field generator, system control unit (SCU), sensor interface unit (SIU), field generator mounting arm and clamp, disposable sensors, six-dimensional reference sensor, six-dimensional palate probe, cables and adapters, and the WaveFront™ Application Software and Documentation.

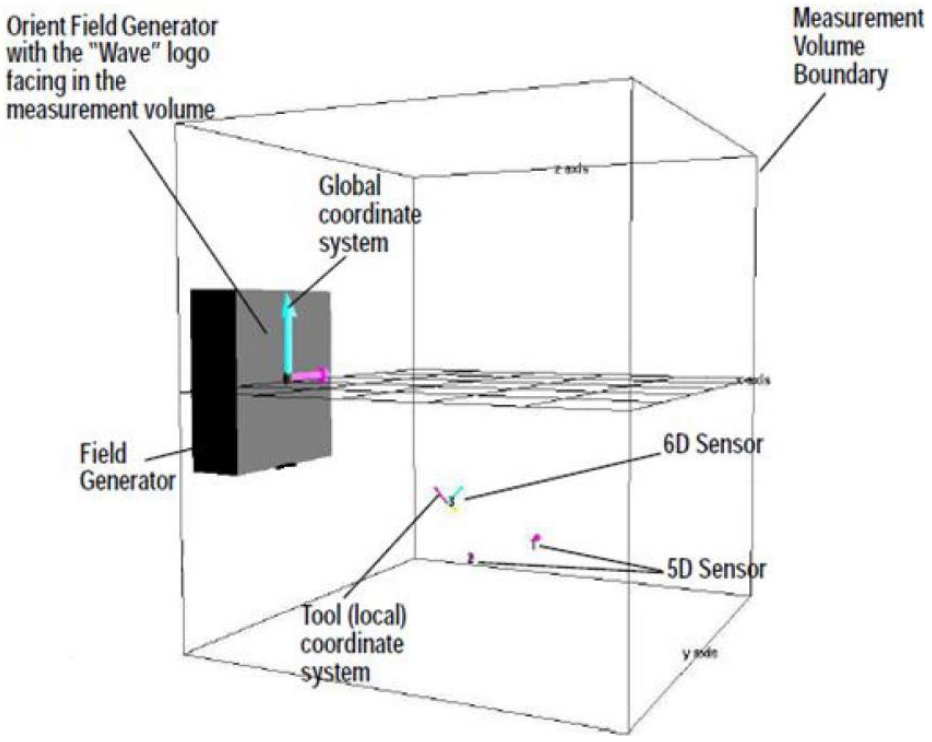

Figure 3: NDI Wave System Field Generator [4]

2.2.2 Experimental Configuration

The experimental set-up for data collection follows the diagram of RASS in Figure 1 and contains the functional decomposition of the software seen in Figure 2. The most basic version of experimental configuration requires the user to attach several wired sensors to the subject's face and then operate the software programs in RASS. WaveFront™, created by NDI and run with the Wave System, is the first software program run and processes the kinematic data gathered by the sensors. A second mapping program is then run to convert the kinematic data into synthesis parameters. VTDemo is then run after the other two programs have completed. The VTDemo program synthesizes the speech based on the inputs from the NDI Wave system and mapping scheme.

For the configuration in RASS that is most commonly used, there are six sensors, five 5-degree-of-freedom sensors and one 6-degree-of-freedom sensor. The purpose of the sensors is to obtain the best model of the vocal tract during speech. Sensor application to human tissues is difficult because tissues do not behave as standard materials. As a result, there are a variety of adhesives used for different parts of the experimental configuration. Stomahesive® Strips are used on the teeth and are similar to double-sided tape. The tongue sensors use small silk patches to increase the surface of adhesion between sensor and tongue. For the lips, a small piece of Super Polygrip® Comfort Seal® Strips is used in combination with glue [7]. The locations of these sensors on a human subject within an electromagnetic field can be seen in Figure 4.

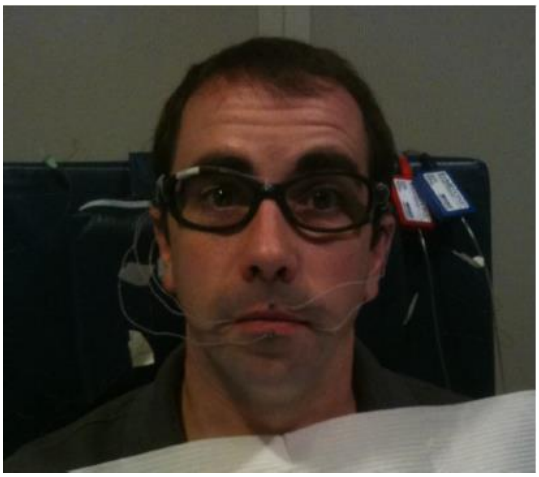

Figure 4: Sensor Placement on Human Subject [8]

A record, which refers to the actual data file of sensor positions, is normally generated in three different forms for each experiment: a bite-plate record, calibration record, and normal record. The placement of sensors is dependent on the type of data collection desired. For example, the bite-plate record is used to correctly orient the subject's personal coordinate system. The x-axis extends across the midsagittal plane

which points away from the front of the subject. The y-axis points upwards, and the z-axis points to the subject's left, which is horizontally perpendicular. The subject additionally wears a pair of glasses with 6-degree-of-freedom sensor attached as a reference sensor. The purpose of these glasses, which are worn for all record types, is to allow for head correction and the shifting of coordinate space.

The creation of a bite-plate record involves placing two sensors on a bite-plate. The first one is placed at the maxillary incisors, and the second is positioned at the bisection between the back molars. A physical bite-plate is created by molding two pieces of softened wax onto a tongue depressor. The subject then bites down on the wax to create a dental impression. Bite marks on the wax allow researchers to correctly orient and place the sensors on the bite-plate. The bite-plate is then re-inserted in the subject's mouth for the duration of the record collection. Sensor positions can be seen in Figure 5 as they relate to the bite-plate [7]. Additionally, the biteplate is shown in a human model in Figure 6.

Figure 5: Bite-Plate with Two Sensor Locations [7]

Figure 6: Bite-Plate in Human Model [4]

The calibration record, which is used to customize the synthesis mapping scheme and normal record, uses five articulatory sensors (seen in Figure 7) plus a reference sensor. The Tongue Blade (TB) and Tongue Dorsum (TD) sensors are placed as close to the midsagittal plane as possible, with the TB sensor closer to the tip of the tongue and the TD sensor toward the back. The Upper Lip (UL) and Lower Lip (LL) sensors are either glued or taped onto the lips. The fifth sensor is then attached to one of the mandibular incisors (MI) with glue and an adhesive strip. Subjects must wear the "orientation" glasses, which hold the reference sensor, during all record collections [7].

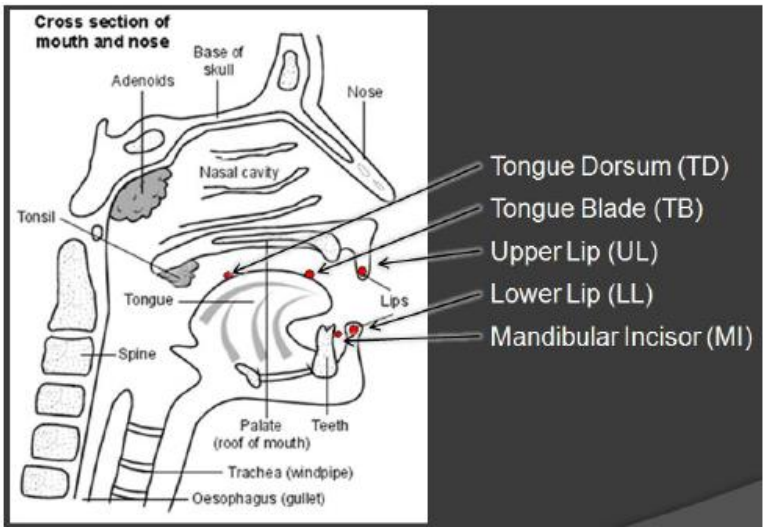

Figure 7: Side View of Sensor Placement [7]

2.3 VTDemo Synthesizer

2.3.1 Source-Filter Model

One of the three main components of RASS, the VTDemo software, is responsible for providing real-time synthesis of the speech signal as the subject's articulatory parameters change. The original creator of the software, Mark Huckvale, from the University College of London, created VTDemo based on the articulatory synthesizer that Shinji Maeda designed called VTCALCS (distributed by Satrajit Ghosh at Boston University). Maeda's program is used to filter a voice signal by developing a vocal tract area function from seven vocal tract parameters. However, VTCALCS does not allow for real-time synthesis in which the effects of changing the articulatory parameters can be audibly detected as they are manipulated. VTDemo extends VTCALCS by incorporating real-time synthesis and other features such as a real-time spectral display, control table for editing dynamic synthesis, and NS and GA parameters for controlling the size of the velopharyngeal port and glottal area, respectively [7] [10].

The source-filter model, which is the basis of VTDemo, is focused on the physical attributes of speech production. More specifically, VTDemo is a cross-sectional, area-driven synthesizer with cross-sections from a very specific physical representation of the vocal tract, based on Maeda's original model. When speech is produced, the excitation wave from the vocal folds passes through the vocal tract and is filtered according to the characteristics of articulators. Since VTDemo is able to control the properties of certain articulators, it is able to adjust the filtering and ultimately the generated speech. A flow diagram as well as a simplified physical vocal tract structure in Figure 8 illustrates the source-filter model [4].

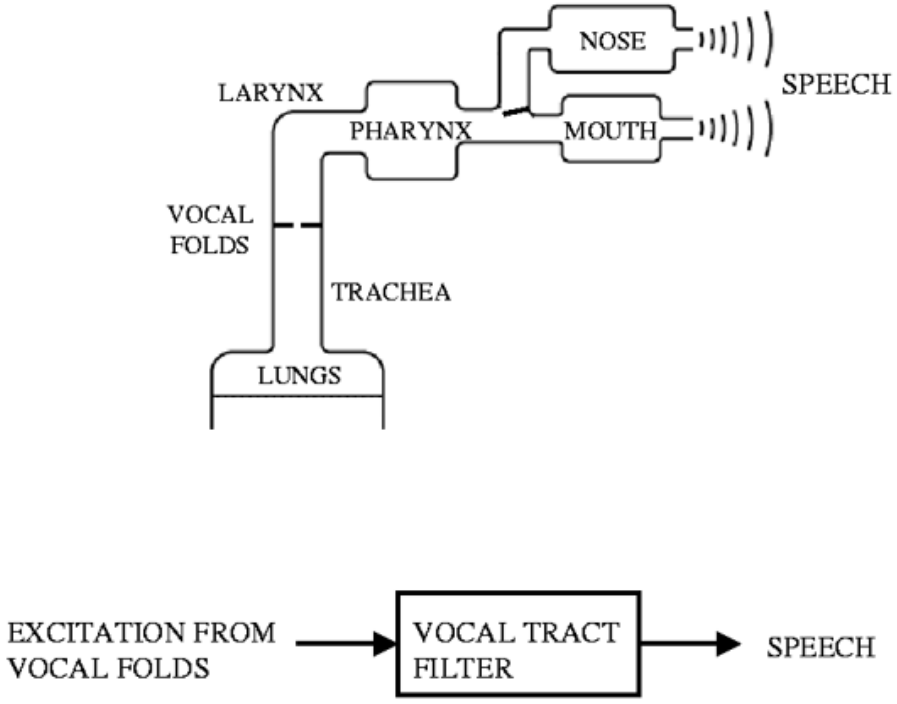

Figure 8: Source-Filter Model [4]

2.3.2 Maeda Model

Shinji Maeda took the source-filter approach and created a vocal tract model with a set of adjustable parameters. These articulatory parameters correlate to the physical positions of different articulators along the vocal tract such as tongue height, jaw position, and lip aperture. Changing the values of the parameters seen in Figure 9 filters the voice and allows for desired speech elements, such as vowels, to be produced.

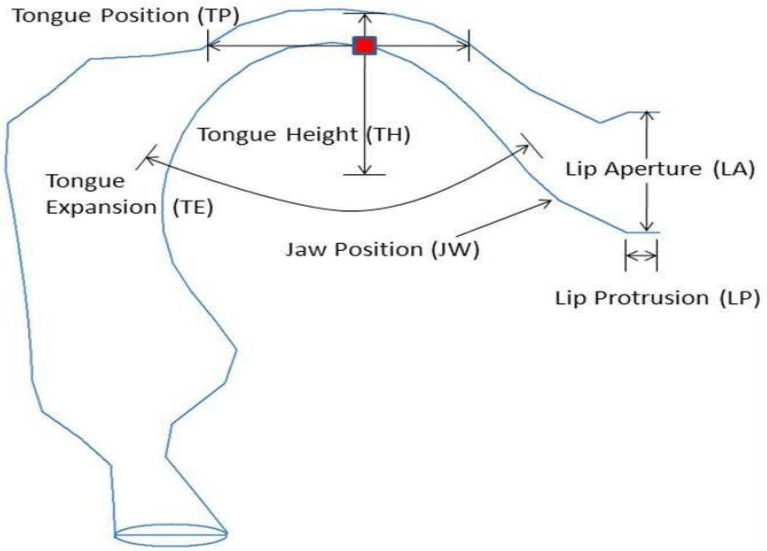

Figure 9: Maeda Model of Vocal Tract [4]

When Maeda developed his model for speech synthesis, the vocal tract shape parameters were a focal point. VTDemo converts seven physical parameters from Maeda's model into a vocal tract area function that filters the incoming voice signal from the source. This filtering mechanism occurs in real-time and is responsible for the final sound of the synthesized speech. The graphical user interface can be seen in Figure 10 [4].

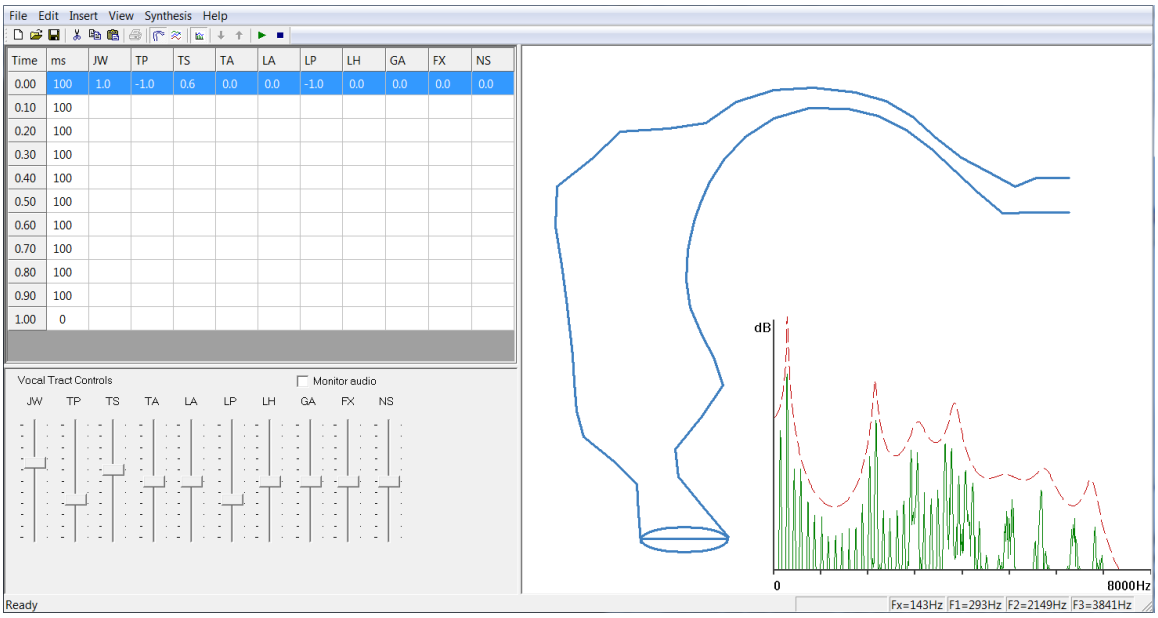

Figure 10: VTDemo Graphical User Interface [10]

The inputs to VTDemo are streamed in real-time from the kinematic data gathered by the EMA system. Sequences of parameter values can also be directly passed to the VTDemo synthesizer through the use of a text file. An example of such a file can be seen in Figure 11, where the vocal tract parameters are displayed in the columns from left to right as appearing in the upper left corner of Figure 10.

```
1  VTPARS2
2  #dynamics=0
3  64.00 1.00 -1.00 0.60 0.00 0.00 -1.00 -3.00 0.00 -4.00 0.00
4  64.00 1.00 -1.00 0.60 0.00 0.00 -1.00 -2.90 0.00 -4.00 0.00
5  64.00 1.00 -1.00 0.60 0.00 0.00 -1.00 -2.80 0.00 -4.00 0.00
6  64.00 1.00 -1.00 0.60 0.00 0.00 -1.00 -2.70 0.00 -4.00 0.00
7  64.00 1.00 -1.00 0.60 0.00 0.00 -1.00 -2.60 0.00 -4.00 0.00
8  64.00 1.00 -1.00 0.60 0.00 0.00 -1.00 -2.50 0.00 -4.00 0.00
9  64.00 1.00 -1.00 0.60 0.00 0.00 -1.00 -2.40 0.00 -4.00 0.00
10 64.00 1.00 -1.00 0.60 0.00 0.00 -1.00 -2.30 0.00 -4.00 0.00
```

Figure 11: VTDemo Sample Input

The standard range of most of the vocal tract control parameters is from -3.0 to 3.0, where the values represent the relative extent of the corresponding Maeda parameters

as shown in Figure 9. The VTDemo program graphically displays an artificial vocal tract model based on these parameter values, as shown in Figure 10. The parameter names and how they related to specific kinematic movement can be seen in Table 1.

Table 1: VTDemo Parameter Description [4] [7]

Parameter	Description	Range	Notes
ms	Segment Duration	n/a	Based on the sampling rate of the NDI Wave system and the speaker; static
JW	Jaw Height	-3.0 to 3.0	Increases with increasing raw value of MIy; dynamic
TP	Tongue Position	-3.0 to 3.0	Increases with decreasing average of TBx and TDx; dynamic
TS	Tongue Shape	-3.0 to 3.0	Increases with increasing average of TBy and TDy; dynamic
TA	Tongue Apex	-3.0 to 3.0	Increases with decreasing average of TBy and TDy; dynamic
LA	Lip Aperture	-3.0 to 3.0	Increasing with increasing distance between UL and LL; dynamic
LP	Lip Protrusion	-3.0 to 3.0	Increases with increasing raw value of LLx; dynamic
LH	Larynx Height	-3.0 to 3.0	Current system: Static, fixed at 0. Proposed system: Set to optimize overall match to target subject
GA	Glottal Aperture	-3.0 to 3.0	-3.0..-2.7 = Open -2.7..-1.5 = Voiceless -1.5..-1.0 = Breathy voice -1.0..1.5 = Normal voice 1.5..3.0 = Creaky voice (**This parameter is not modified currently – set at 0**) static
FX	Fundamental Frequency (F0)	-3.0 to 3.0	Adult Male: 89-191Hz Adult Female: 161-299Hz Child: 199-361Hz Current system: Static, fixed at 0 Proposed system: Real-time adjustments to match to target subject
NS	Velo-pharyngeal Port (Nasality)	0.0 to 3.0	**(This parameter is not modified currently – set at 0)** static

In the table above, there a few terms that require clarification. The abbreviations MI, TB, TD, UL, and LL stand for the following midsagittal placements, respectively: middle lower incisor, tongue blade (5 mm behind tip), tongue dorsum (40 mm back), upper lip, and lower lip. The placements are often listed with an "x" or "y" following the abbreviation which stands for the orientation in 3D space. Additionally, the laryngeal height parameter refers to the length of the larynx, which is the bottom of the vocal tract model in Figure 10. The glottal aperture parameter, which constricts the airflow into the model, refers to the size of the glottal opening at the very bottom of the vocal tract model. Finally, the nasality parameter quantifies the size of the opening to the nasal cavity and can be controlled in real time.

The set of parameters in Table 1 can be further described as kinematic, structural, and excitation parameters. The kinematic parameters (JW, TP, TS, TA, LA, LP, and NS) are time-dependent and control the structure of the vocal tract based on articulatory movements. Excitation parameters (FX and GA), which represent acoustic characteristics of a speaker, change over time as well but can't be controlled by kinematic motion. The structural parameters (LH and SF) are time-independent and correspond to a fixed part of the vocal tract model. SF stands for scaling factor and is defined in Section 3.1.1. Each speaker is assigned one set of these structural values for the experiments. In this thesis, the methods of time-independent (fixed) parameter determination refer to the structural parameters, and time-dependent parameter determination refers to the FX excitation parameter.

During synthesis, the VTDemo software uses a low-order linear predictive coding (LPC) analysis to represent the spectral envelope of the speech signal which allows the

current first, second, and third formant values in Hz to be displayed. These values fluctuate as a result of fluid articulatory parameters and can be seen in the lower right hand corner of Figure 10 along with the associated spectrum. The formant values are useful for researchers to quantify the impact of real-time synthesis parameter adjustments.

2.3.3 Modified VTDemo

In Marquette's Speech and Swallowing Lab, researchers made some minor changes to the VTDemo graphical user interface in order to better match the vocal tract model to the subjects. One change is the slider for the FX parameter, which now ranges from -6.0 to 6.0 instead of -3.0 to 3.0. The FX value can also be directly specified in Hz by entering a value in the fundamental frequency text box. These two different controls for fundamental frequency are reconciled in that every increment of 1.0 on the slider moves the value up or down 17 Hz, with the number in the text box serving as the zero-point for FX. Before the text box was available, the fundamental frequency was set at the defaults of 140 Hz, 230 Hz, and 270 Hz for a male, female, and child, respectively. There is also an added scaling factor textbox, which allows the size of the vocal tract to be adjusted using the scaling factor variable, SF. Finally, there is a "connect" and "disconnect" button at the top of the graphical user interface (GUI) that allows the RASS configurations such as biteplate and sensor data to be loaded. These changes are seen in Figure 12.

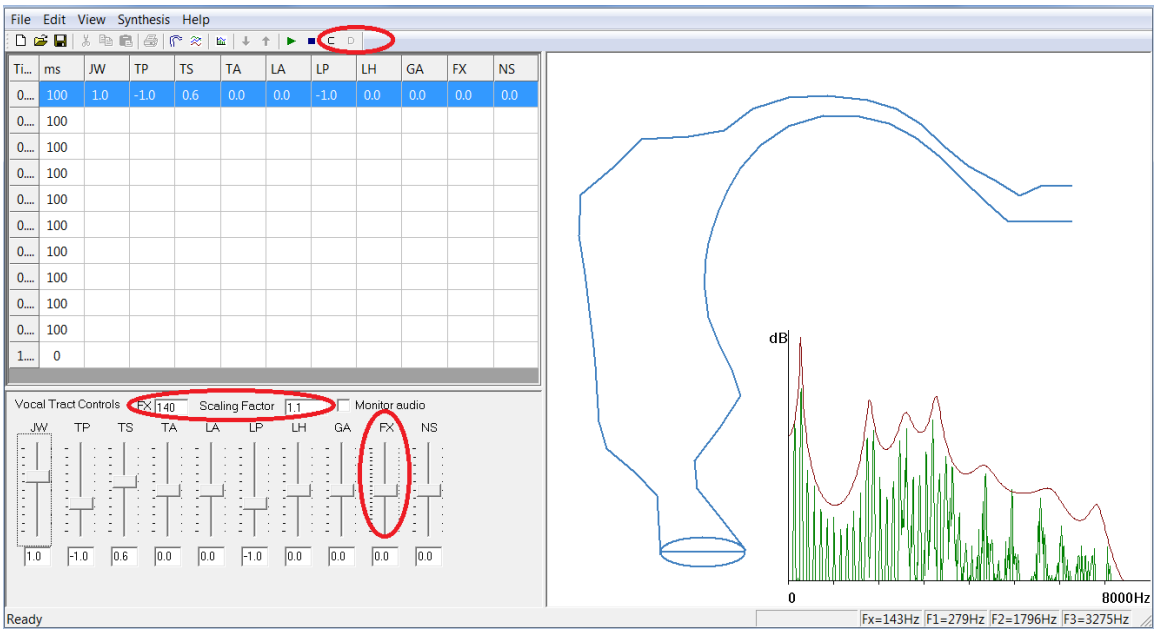

Figure 12: Updated VTDemo GUI

2.4 Kinematic Mapping to Synthesis Parameters

In order to bridge the gap between the NDI Wave System and VTDemo, a mapping scheme is employed to convert the kinematic data into synthesis parameters. There are two mapping systems trialed in RASS, linear interpolation and quantile analysis. However, the quantile method, which looks at the distribution of quantiles across a range, is more significant for subject and synthesized speech comparisons because it is stable and used in the majority of experiments in Marquette's Speech and Swallowing Lab. This mapping method breaks the sensor data values into 61 quantiles and then maps those quantiles to synthesis parameter values based on their breakpoints (where one quantile of sensor data ends and the next one begins). In order to start the calibration for mapping, the speaker is required to read "The Caterpillar" passage, from which RASS determines kinematic sensor dynamic range and position distributions and establishes the mappings between these and the synthesis parameters. The passage,

which requires a wide range of articulatory movements, is relatively easy to pronounce and provides a large span of sensor data to better characterize the subject's speech [11].

The 61-point quantile mapping method uses the passage to create a subject-specific calibration matrix with 8 columns (and 61 rows). This term, "calibration matrix," is a slight misnomer since it serves as a mapping matrix to convert kinematic data to synthesis parameters. The first column contains the breakpoints for the full-range of synthesis parameters (breakpoints range from -3.0 to 3.0 by 0.1 increments), and the last column contains the breakpoints for the half-range of synthesis parameters (breakpoints range from -1.5 to 3.0 by 0.075 increments). The middle six columns hold the breakpoints for each of the individual sensor data variables (columns 2 and 4 are reverse-ordered because the mapping is inverted). The sensor breakpoints, which are the bounds by which to numerically sort sensor data, are calculated by dividing each of the sensor columns into 61 quantiles based on the individual sensor values. An example matrix is displayed in Table 14, Appendix A.

Once the calibration matrix is filled, the quantile method can appropriately map a subject's kinematic data to synthesis parameters for a desired segment of speech. The relevant sensor data variables for each synthesis parameter, as described previously in Table 1, are compared to their respective sensor columns in the calibration matrix to find which two breakpoints they are between. A VTDemo synthesis parameter value at a specific time is calculated by linearly interpolating between the two appropriate VTDemo parameter breakpoints based on the specific kinematic data sensor value and its location between kinematic data breakpoints. The overall goal is to map each kinematic sensor

data value to a specific VTDemo parameter. A full description of this process and other mapping methods in RASS are detailed in Zhou's thesis [4].

2.5 Audapt System

After speech is synthesized by the VTDemo software, there is an optional component for speech processing to enhance auditory feedback. Acoustic perturbation software, called Audapt, can alter the speech signal in order to advance specific learning outcomes [12]. For example, shifting the F1 and F2 values of an acoustic signal before it is fed back to the subject can elicit involuntary changes in the movements of a subject's articulators and ultimately change the subject's pronunciation. Additionally, Audapt can modify the acoustic signal through the use of pink noise masking auditory feedback, alternating stimuli, and adjusting levels of perturbation. The purpose of Audapt is to provoke involuntary learning and ultimately change the way the subject speaks [13].

2.6 Involuntary Learning

Sensorimotor adaptation (SA) is an involuntary learning process that has the potential to be used for speech rehabilitation applications. It can be understood as the neural response to perturbed auditory feedback. Within the human body, sensory feedback systems exist for functions such as movement of the limbs and speech production. If the desired outcome is not reached, the feedback loop enables the body to correct the motions to achieve the result. Over a period of time, the compensation resulting from this feedback loop becomes learned motor behavior. This behavioral change has been observed during studies of upper limb movement, in which visual

feedback served to modify the brain's prediction of the result of limb movement. The result was the recalibration of the limb's motor behavior to attain the desired outcome [3].

During studies involving SA, the compensatory motor response to sensory feedback is monitored while the sensory feedback is perturbed or masked. It has been used for a variety of neurorehabilitation applications pertaining to effects of brain injury such as hemineglect, gait, and upper limb movement. Often virtual reality environments are used as the patient interfaces for SA rehabilitation systems. The potential of SA phenomena is currently being investigated with regard to speech neurorehabilitation [14]. Speech SA is characterized by the modification of articulator movements as a result of sensory-feedback perturbation. The perturbation commonly consists of a shift in formants, which results in a compensatory motor response. When the compensatory movement persists when the feedback is masked, the subject is considered to demonstrate adaptation [3].

Previously, SA has not been studied in individuals with severe motor speech disorders because the established techniques for speech adaptation, such as those used in Audapt, require acoustically high-quality speech from the subject. The commonly used linear predictive coding-based approaches do not function correctly with the speech produced by those with severe motor speech disorders and often do not accurately resynthesize the speech of those without speech disorders [14].

In order to effectively use SA phenomena with a wider variety of individuals, Dr. Jeffrey Berry and a team of researchers, proposed a speech adaptation technique that does not require acoustic resynthesis of the subject's speech [14]. Instead, an articulatory

speech synthesizer utilizes data produced by an electromagnetic articulography (EMA) system. The focus of this proposal was the RASS system as previously discussed. The researchers suggested that clinical benefits of the NDI Wave system included automated head movement correction and average sensor tracking errors less than 0.5 mm when the sensors moved at velocities in the upper range of typical human speech kinematics. The acoustic output of the speech synthesizer is received by the subject and utilized as auditory feedback, which can then be perturbed using an established, acoustic-based method. As this EMA method utilizes the movements of the articulators as parameters for speech resynthesis, the sound quality of the subject's speech does not matter, allowing this method to be used with subjects who suffer from severe motor speech disorders [3] [14].

This technique for involuntary learning was used in a study with five human participants, where five phases existed for each subject. The first phase was characterized by no perturbation in the auditory feedback. The second was a ramp phase in which gradual perturbation occurred through increasing shifts in the first and second formants. The third phase occurred at the stage in which the maximum perturbation in the first and second formants occurred. The fourth was a masking phase, in which auditory feedback was masked with noise. The fifth and final phase was a return phase in which the formants were gradually decreased to their original values. Three of the five subjects experienced significant shifts to compensate for the changing formants, which is consistent with the principles of SA [14]. Although the results of auditory feedback from an SA system can be influenced by coarticulatory context, it is suggested that the use of

EMA with an articulatory speech synthesizer can take advantage of SA phenomena for rehabilitation in subjects both with and without severe motor speech disorders [3] [14].

2.7 Summary of Background

The RASS system is designed to synthesize speech based on the articulatory movements of a human subject. More specifically, the system consists of an electromagnetic articulography component which tracks the movements of a subject's articulators during speech. The gathered kinematic data are sent into a mapping algorithm to convert the sensor positions into synthesis parameters. Those parameters are streamed into the VTDemo component of RASS for real-time speech synthesis. Once synthesized, a software program called Audapt is used to adjust the synthesized speech before it is fed back to the subject. Audapt can be used to control learning objectives and elicit a change in a subject's fine motor behavior when the adjusted speech is heard through headphones. An understanding of the RASS system is necessary to determine how to best match the synthesis parameters to the subject's acoustic characteristics in order to increase involuntary learning.

CHAPTER 3: TIME-INDEPENDENT PARAMETER MATCHING

3.1 Background

3.1.1 Time-Independent Parameters Under Investigation

There are three main components to RASS as previously described in Chapter 2: the EMA system, mapping algorithm, and VTDemo software. Table 1 provided a detailed description of the VTDemo synthesis parameters, most of which are derived from the subject's kinematic data, that control the vocal tract model. The non-articulatory-based parameters that can be used to increase the match between individual subjects and the synthesizer are LH, SF, FX, and GA. While all can be adjusted over time through VTDemo, the first two of these, LH and SF, relate to the physical size of the vocal tract, and therefore, it is reasonable to assume that they should be fixed for a specific subject. The latter two of these relate to vocal characteristics which are associated with changing speech characteristics and are therefore treated as time-dependent. Of the four, the focus is on the time-independent parameters LH and SF and the time-dependent parameter FX, related to pitch. GA (glottal aperture), while a potentially impactful parameter, is not the focus of this thesis.

Laryngeal height (LH) controls the length of the modeled vocal tract's larynx and can be adjusted from -3.0 to 3.0 by 0.1 increments. A small LH parameter value represents a short larynx, and a large value intuitively represents a longer larynx. Scaling factor (SF) is a new parameter developed by Marquette's Speech and Swallowing Lab which allows users to linearly scale the overall size of the vocal tract by either shrinking or expanding the model. A low scaling factor, such as 0.8, decreases the size of the vocal tract model to 80% of its original size, while a factor of 1.1 increases the size of the vocal

tract to 110%. The goal of this chapter is to determine the effectiveness of altering the LH and SF parameters to increase the match between synthesized speech and subjects' acoustic characteristics.

3.1.2 Introduction to Time-Independent Parameter Determination Methods

Laryngeal height and scaling factor parameters have the potential to improve the similarity between synthesized speech and a subject's speech. However, while the two parameters are easily accessible for subject matching adjustments, the present use of VTDemo during experiments does not change the default SF and LH parameters for speech synthesis. Since the SF and LH parameters control the physical characteristics (overall size and laryngeal length, respectively) of the vocal tract, they could be used to better customize experiments for subjects who widely differ in such physical attributes, thus improving the subject-matching goal of the RASS system.

One way to determine the most appropriate SF parameter for a subject is to create a formant look-up table for vowels synthesized at different scaling factors and choose SF in order to most closely match the synthesis model's formants to the subject's acoustically measured formants. This table focuses on the first and second formant values for specific pre-selected artificial vowels as well as the range of scaling factors from 0.8 to 1.3 as performed in previous experiments in Marquette's Speech and Swallowing Lab. A more detailed explanation of the look-up table used in this thesis can be found in Section 3.2.2. Vowel spaces, which are a subject-specific characterization of vowels plotted in the formant space, can be used to distinguish subjects. While the plot in Figure 13 shows a vowel space containing nine common vowels, the focus of the look-

up table is on the four corner vowels in the figure denoted by the words “heed,” “had,” “hod,” and “who’d.” These corner vowels maintain the general outline of the vowel space while reducing the potential complexity of the look-up table.

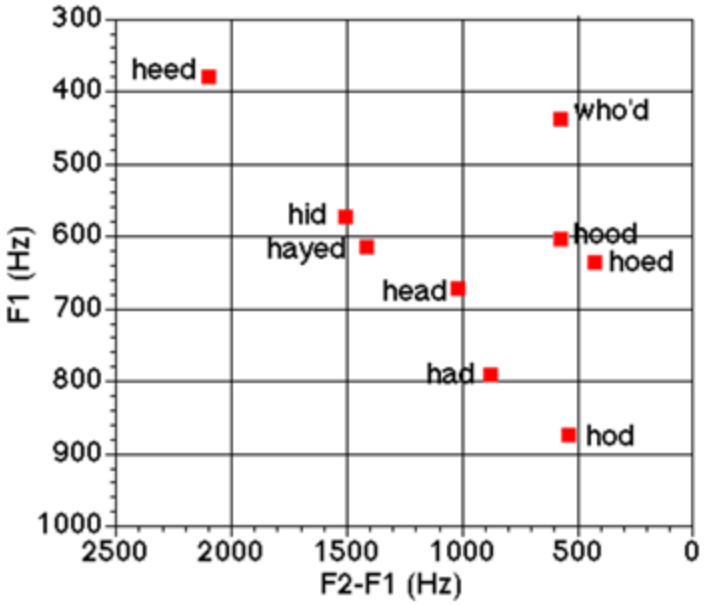

Figure 13: Nine Common Vowels Plotted in Human Vowel Space [15]

There are two methods that have been implemented with the current RASS system for selecting the best SF synthesis parameter based on vowel formant value comparisons between subject vowel measures and corresponding synthesized vowels. Both of these methods are based on the same concept, which is to experimentally determine the best synthesis parameters by comparing the formant values of a selected set of vowels with varying parameters against a speaker’s acoustically measured formant values. Each of the two methods has a different mechanism for measuring the difference between a set of synthesized and speaker vowel formants. The first of these is based on the minimum sum Euclidean distance (SED) in the formant space of selected corner vowels, while the second is based on the maximum vowel space overlap (VSO) created

by those same corner vowels. In both of these approaches, a look-up table that lists measured vowel formant values for specific synthesis parameters as a function of scaling factor is used to identify the scaling factor giving the lowest metric. For the vowel space overlap method, the synthesized vowels are plotted in the vowel space at each increment of the scaling factor (0.8 to 1.3 by increments of 0.02). Connecting the points between the same increments of scaling factors for each of the corner vowels creates a polygon. An example of the smallest and largest polygons (SF set at 1.3 and 0.8, respectively) based on the 4 corner vowels is shown in Figure 14.

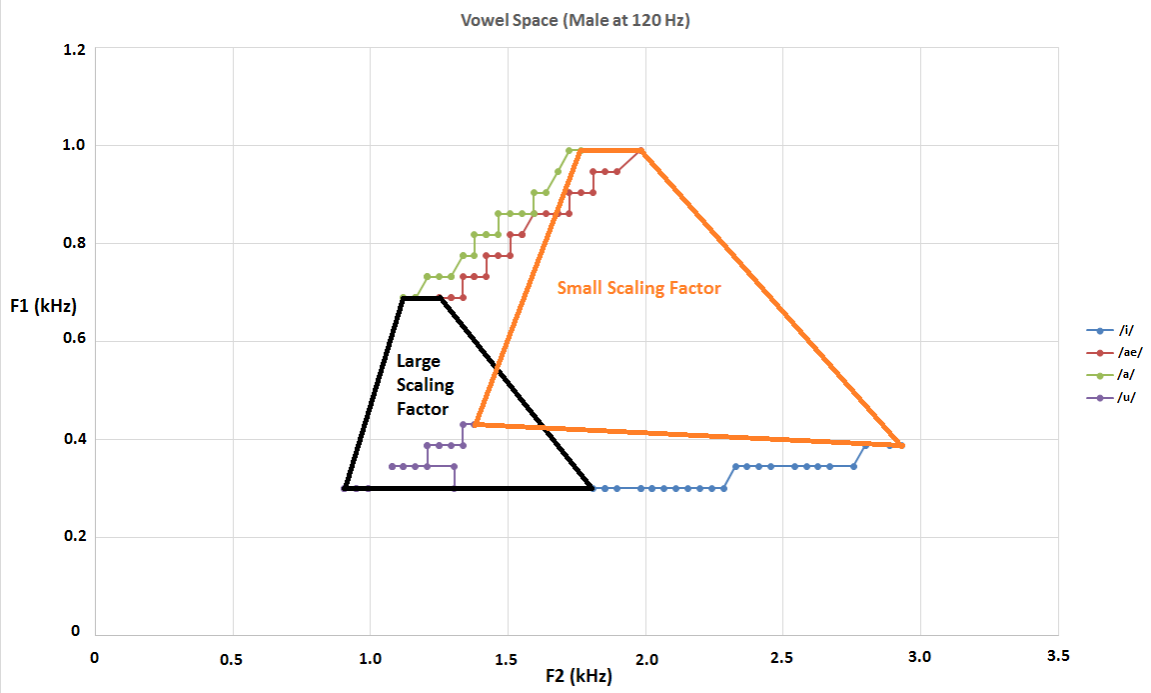

Figure 14: Vowel Space Scaling Factors

When the overlap algorithm is run, a polygon is calculated for each increment of possible scaling factors. The area of each polygon is compared to the area of the subject's polygon, which is created by plotting the subject's corner vowels in the vowel space. The goal is to choose the scaling factor which enables the maximum amount of

area from the subject's polygon to overlap with the synthesized polygon. The overlap indicates similarity between the vowel spaces and ultimately a better match between synthesis parameter values and the subject's acoustic characteristics. In order to simplify the overlap model, Marquette's Speech and Swallowing lab changed the number of corner vowels from four to three since the /ae/ and /a/ vowels are located close to each other in the vowel space. The three vowels currently used are /i/, /a/, and /u/ ("heed," "hod," and "who'd" as seen respectively in Figure 13).

Figure 15 displays a vowel space overlap plot using three corner vowels, where the blue polygon represents the subject's formant values and the green polygon represents the synthesized formant values that provide the greatest overlap with the subject. In this example, a scaling factor of 1.1 was used to generate the polygon because that value resulted in the highest degree of overlap with the subject's polygon.

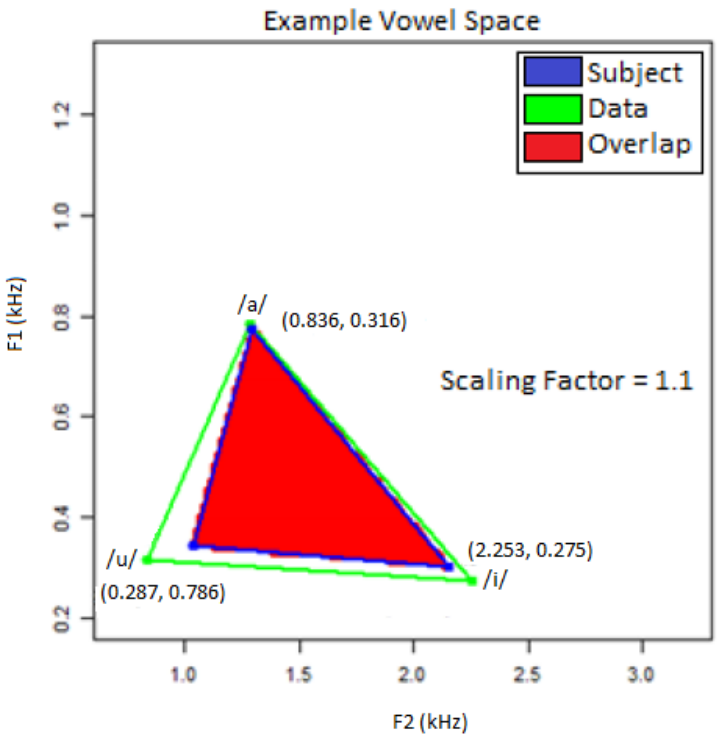

Figure 15: Vowel Space Overlap

One potential weakness of the two above methods of synthesis parameter to subject matching is that they only consider a single variable, scaling factor (SF), using a table that was constructed with a fixed fundamental frequency (F0) parameter (FX) and a fixed laryngeal height parameter (LH). Laryngeal height, while also affecting formant values, is a parameter representing a physical characteristic that varies by subject. Thus, there could be a better set of parameters that characterize the subject than those chosen by the current methods. As evidenced by the plot in Figure 16, there are many instances where different synthesis parameter combinations lead to similar F1 and F2 values. This many-to-one characteristic leads to a disparity in synthesis parameters for close points, which ultimately changes the way speech is synthesized in VTDemo. Since the speech synthesizer is not able to perfectly produce formant values that match the subject's, the potential for inaccurately choosing parameters is magnified by the close proximity of points as seen in Figure 16. It would be preferable to identify a method that considers all synthesis parameters in a way that is representative of the VTDemo synthesis model, i.e. finding an optimal fixed value for both LH and SF while varying the F0 value to match the subject's F0.

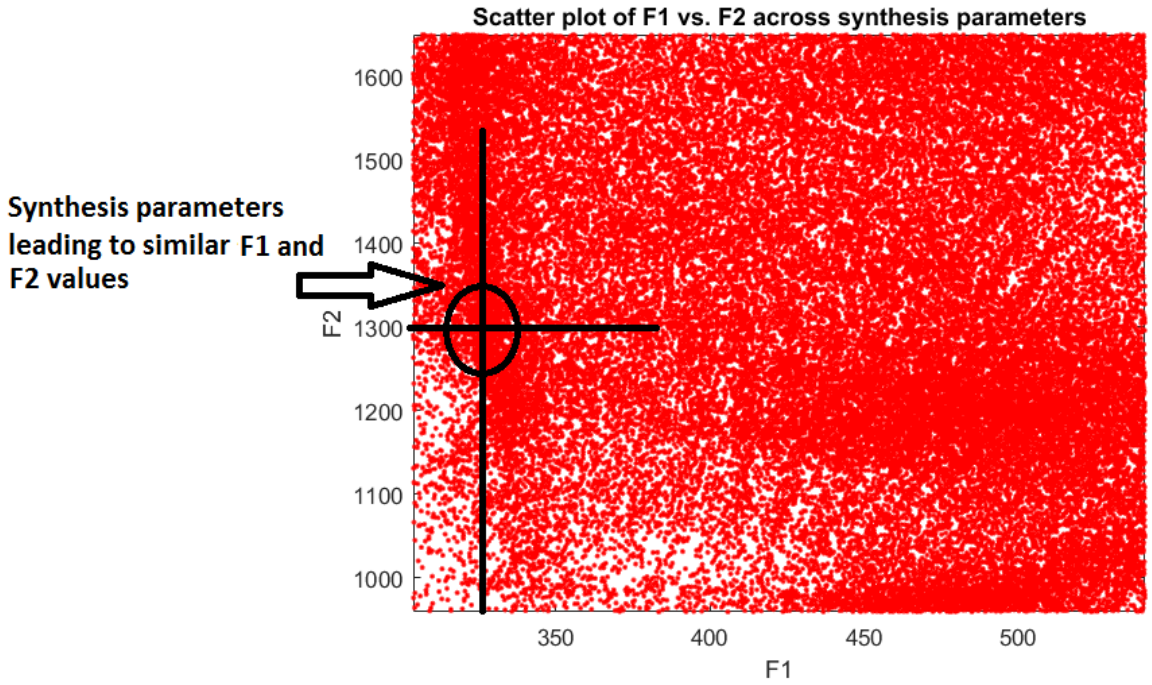

Figure 16: Distribution of Vowel Formants across VTDemo Synthesis Parameters [4]

The laryngeal height parameter is currently unused as a synthesis parameter in VTDemo and corresponds to the length of the larynx in the model. Although this parameter isn't measured by sensors, it can be implemented into the subject-matching algorithms by expanding the current look-up table with pre-synthesized formant values for different LH values. This expansion would allow researchers another parameter to further specify the set of synthesis parameters that most closely align the synthesized formants to the subject's formants for the corner vowels. The determination of the laryngeal height parameter would be performed in tandem with choosing an appropriate scaling factor, both time-independent parameter settings, in order to best match the synthesized voice to the subject's acoustic characteristics.

In order to investigate this question, the formant-matching methods used for determining scaling factor will be expanded to look at the impact of varying FX and LH as well.

3.2 Investigation of Relationship between F0, SF, LH, and Both F1 and F2 for Subject Matching

3.2.1 Determining Effect of LH and SF on F1 in Subsample

Time-independent synthesis parameters, specifically SF and LH, have the potential to affect synthesized formants in VTDemo. With the ultimate goal of determining synthesis parameters that best represent a subject's acoustic characteristics, a relatively small experiment was devised to investigate the effects of varying LH and SF during speech synthesis. This experiment was significant because it provides the justification for expanding the look-up table previously mentioned in Section 3.1 to include the LH parameter. Using the default VTDemo synthesis parameters for one of the three corner vowels, /i/, a ".vtd" file was created, which is an input file for the speech synthesizer. This file was then used as the basis to construct larger input file that looped the synthesis parameters for /i/, increasing the LH parameter from -3.0 to 3.0 by 0.1 every 65 ms.

Each time the synthesis experiment was run, the scaling factor was incremented by 0.02, resulting in 26 total trials that spanned the entire range of the LH parameter and SF parameter (0.8 to 1.3). During the trials, the first formant values as estimated by the VTDemo program were tracked and stored in a text file for analysis. The aim of this experimental configuration was to analyze the relationship between the scaling factor and laryngeal height parameters as they applied to the first formant value. Figure 17 shows the plotted data points for /i/ with the previously mentioned combinations of LH and SF parameters in a three-dimensional view. Figures 18 and 19 present specific angles from the three-dimensional plot that are more revealing of trends.

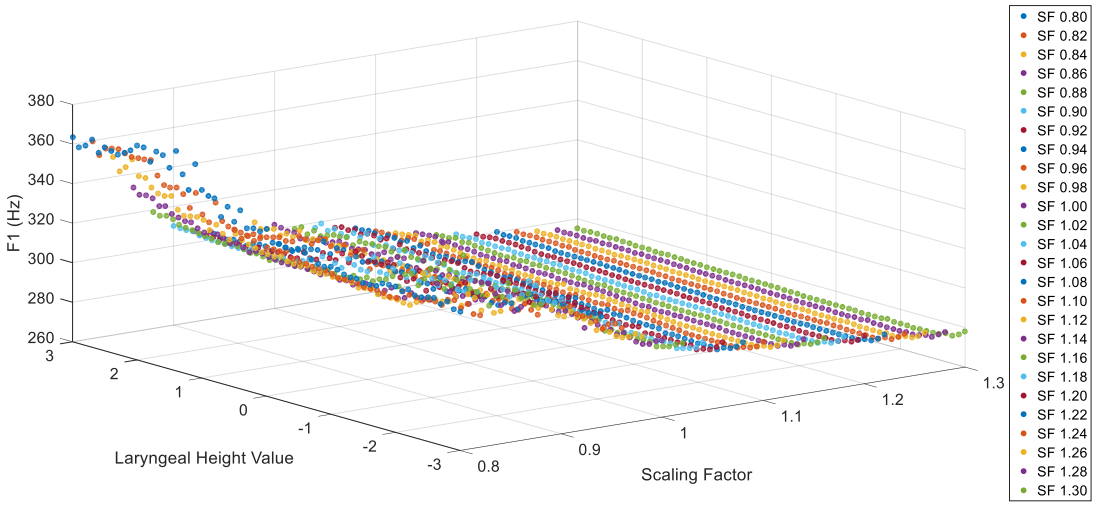

Figure 17: Relationship between Laryngeal Height, Scaling Factor, and F1 for /i/ (with SF legend)

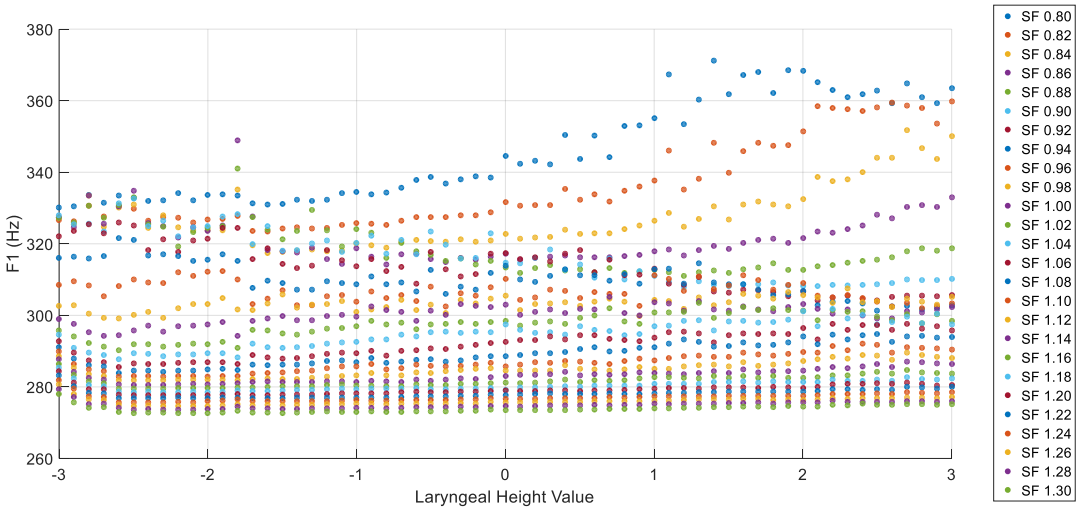

Figure 18: F1 vs. Laryngeal Height Value from Figure 17 (with SF legend)

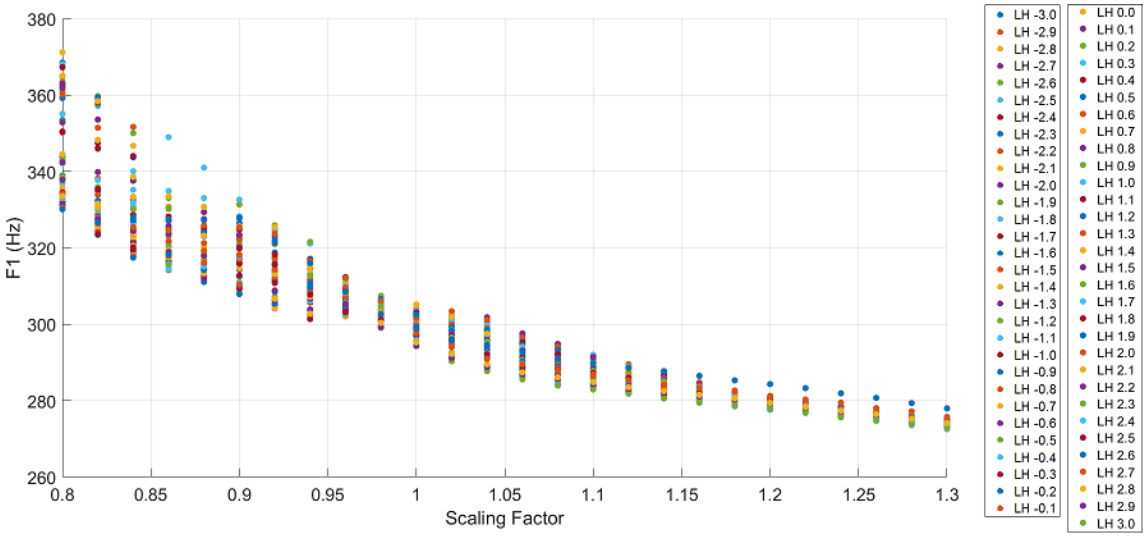

Figure 19: F1 vs. Scaling Factor from Figure 17 (with LH legend)

Evaluating the plots in Figures 18 and 19 reveals that low scaling factors (smaller vocal tract) are more significantly affected by LH parameters than high scaling factors. There is also a larger variance among the data points in Figure 18 at high LH values than low LH values. These observations support the idea that the LH parameter plays a significant role in affecting the first formant of /i/. However, as seen by the non-linear arrangement of data points for each set scaling factors in Figure 18, the relationship between LH, SF, and F1 is difficult to predict. In Figure 19, the set of LH parameters is elongated along the y-axis range (F1) for lower scaling factors and condensed for higher scaling factors. This trend again shows that the LH and SF parameters interact the most when the LH is high and the SF is low. The interaction between these two time-independent parameters allows for multiple combinations of LH and SF that produce similar F1 values. Since the goal of RASS is to match the synthesis parameters to a subject's acoustic characteristics, these multiple combinations of LH and SF for a target

F1 value allow for a greater resolution in the matching algorithms (Section 3.1) compared to solely using the SF parameter.

The addition of an LH parameter aids in creating more unique combinations of synthesis parameters but still does not allow for exact one-to-one relationships between the synthesis parameter settings and first formant value. A vowel can have several distinguishable formants; however, the clarity and accuracy of the peaks decrease as the formant number increases [16]. Based on the number of similar SF and LH combinations that produce the same F1 value in the two figures above, researchers could still benefit from the use of a second formant value to distinguish LH and SF combinations from each other and create another criterion for subject-matching.

3.2.2 Methodology for Full Investigation of Time-Independent Parameters

In order to more accurately determine synthesis parameters that correspond to a subject's vocal tract characteristics, an expanded set of formant data collected from a larger array of synthesis parameters was needed. While the initial data collection, which included LH, SF, and F1 values, provided a useful analysis of trends between the time-independent parameters under investigation, a second formant value recorded over more vowels and F0 values was required to create a comprehensive look-up table. The idea of this data collection was to synthesize the three corner vowels at every combination of scaling factors (0.8 to 1.3 by 0.02 increments), laryngeal height parameters (-3.0 to 3.0 by 0.1 increments), and F0 (-4.0 to 10.0 by 0.1 increments). Note that FX, while a time-dependent parameter, can be set to fixed values to increase the size of the look-up table. For the vowels /i/, /a/, and /u/, the settings for the synthesis parameters JW, TP, TS, TA, LA, and LP were [1.0, -1.0, 0.6, 0.0, 0.0, -1.0], [-1.8, 2.7, -1.8, 0.0, 0.3, 0.0], and [3.0,

1.9, 1.6, -0.3, -0.6, -0.2], respectively. These parameter combinations were determined by the preset vowel settings built into the VTDemo software.

During the data collection, the default F0 value corresponding to FX = 0.0 was set to 140 Hz. This created a dynamic F0 range of 72 Hz to 310 Hz controllable via the FX parameter, which covers the normal range of human fundamental frequencies [17]. As seen in Table 2, integer values of the FX parameter represent increments of 17 Hz. The 0.1 increments used in creating the table thus represent F0 increments of 1.7 Hz. The look-up table serves as a large database for matching a subject's first and second formant values to synthesized ones, and is constructed based on combinations of F0 (FX), scaling factor, and laryngeal height parameters.

Table 2: FX Values for Fundamental Frequency

FX value	F0 (Hz)
-4.0	72
-3.0	89
-2.0	106
-1.0	123
0.0	140
1.0	157
2.0	174
3.0	191
4.0	208
5.0	225
6.0	242
7.0	259
8.0	276
9.0	293
10.0	310

A preliminary step in synthesizing formants for the look-up table was to produce input files for VTDemo that cycled through the different combinations of LH and FX

parameters for each of the three corner vowels. For each vowel, there were 26 different scaling factors tested (0.8 to 1.3 by 0.02 increments), 141 different F0 values (-4.0 to 10.0 by 0.1 increments), and 61 different LH values (-3.0 to 3.0 by 0.1 increments) representing a total of 223,626 distinct table values.

It is important to note that this table is composed of vowel formant values synthesized by preset artificial kinematic parameters which do not exactly match a particular speaker. Rather, they were gathered from the VTDemo program as the default vowel settings of the Maeda synthesizer. Ideally, a unique look-up table would be generated for each individual speaker using the subject's actual kinematic parameter values to produce subject-specific formant values for the vowels. However, for the purpose of designing a tool that can be used uniformly across subjects and the large amount of time to generate a look-up table on a subject-by-subject basis, the artificial preset parameters were selected for this work.

3.3 Characterization of Speaker Similarity

3.3.1 Methods for Characterizing the Match between Speaker and Synthesized Speech

There are several ways to analyze the formant data synthesized by the VTDemo software in Section 3.2. Since the goal of the research is overall synthesis-parameter-to-subject matching, methods that minimize error between the subject and synthesizer are a logical choice. The Euclidean distance formula, as seen in Equation 3.3.1, was originally used with the RASS before the addition of the expanded look-up table with the LH and FX parameters. The formant values chosen as the best match to the subject in the new look-up table now represent a scaling factor and a laryngeal height parameter.

$$\text{Euclidean Distance} = \sqrt{(F_{2Subject} - F_{2Synthesized})^2 + (F_{1Subject} - F_{1Synthesized})^2} \qquad (3.3.1)$$

In order to re-evaluate the Euclidean distance method, a MATLAB script was written that prompts the user to input the subject's F1 and F2 values (in Hz) for the three corner vowels. The script steps through the look-up table for each corner vowel, comparing the Euclidean distance between synthesized and subject formant values as shown in Equation 1 above. The Euclidean distances for each combination of SF, LH, and FX are stored in an array for each corner vowel. To find the LH and SF parameter values that minimize the error between the subject and synthesizer, the Euclidean distances from corresponding combinations among the three vowels are added together and put into a new array. The minimum value in the array containing the sum of Euclidean distances across the three corner vowels signifies the best match to the subject's acoustic characteristics. LH and SF are extracted from the look-up table row associated with the smallest sum of Euclidean distances.

Another way to find the best match of scaling factor and laryngeal height parameter values is through the Vowel-Space-Overlap method previously mentioned. Again, this method was altered to better fit the expanded look-up table. A MATLAB script was written that prompts the user to enter the subject's first and second formant values for the corner vowels and then loads the text files containing the corner vowels' synthesized formant values. The first and second formant values for the subject and synthesized data are then assigned to y-coordinate and x-coordinate arrays respectively. For the length of the synthesized formant array, the coordinates of the synthesized vowels in each row are plotted in the vowel space (F1 vs. F2) as well as the subject's

coordinates. Connecting the synthesized data points between the corner vowels for each of the LH, SF, and FX combinations forms numerous polygons in the vowel space. The area of each polygon is then calculated and recorded. The same area calculation is performed on the subject's polygon in order to compare the two sets of data. At any one time, there are only two polygons drawn in the vowel space, the subject's polygon and one of the synthesized combinations. A MATLAB function written to calculate the overlap between the subject and synthesized polygons is run, and that overlap area is stored in an array. The process is repeated until there is an array of overlap areas for each combination of synthesis parameters used to generate the formant values for the three corner vowels.

A maximum overlap area signifies the highest degree of similarity between the subject's vocal tract characteristics and the synthesized parameters. Ideally, the overlap would be 100%, however that is highly unlikely given the significant number of formant combinations. Once the maximum overlap is chosen from the array, the corresponding LH and SF parameters are extracted and output to the user. Additionally, the vowel space overlap method displays the appropriate plot that results in the maximum overlap as seen in the example in Figure 20. In this figure, the subject (blue) and synthesized formant values (red) triangles are plotted in the vowel space with the green portion equivalent to their overlapping areas. Subject 27 is one of the illustrative subjects studied later in Section 3.4 of this thesis.

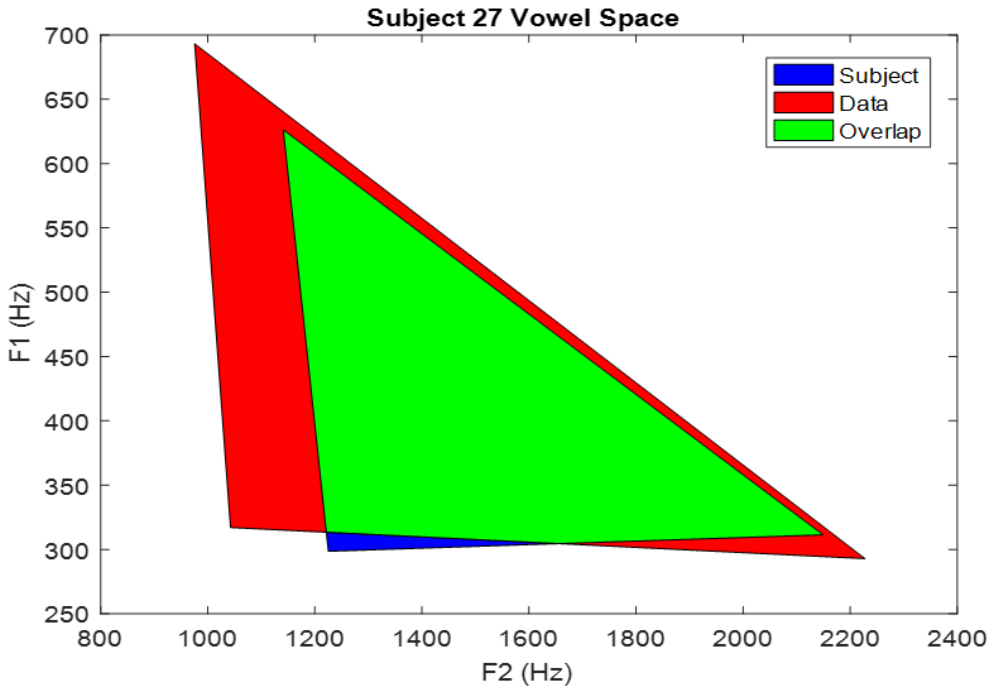

Figure 20: Vowel-Space-Overlap Method

3.3.2 Method Evaluations

Both the Sum-Euclidean-Distance and Vowel-Space-Overlap methods were designed to determine synthesis parameters that best represent a subject's acoustic characteristics. The effectiveness of the two methods' representation of acoustic characteristics was investigated by analyzing their potential weaknesses. While both approaches are dependent on the accuracy the VTDemo synthesizer's formant values in the look-up table, the overlap method contains an additional area of concern. The algorithm is designed to find the greatest possible overlap area between the subject and synthesized parameter corner vowel triangles in the vowel space. Due to the size of the look-up table, it is possible in some experiments that a combination may exist where the synthesized triangle completely surrounds the subject's triangle. While the overlap is 100% in this case, the corners of the synthesized triangle may not be best matched to

synthesis parameter values. An example of this inaccuracy can be seen in Figure 21, where the red triangle completely encompasses the blue triangle and has vertices located at large distances from the blue triangle's vertices. Future work could be performed to improve the algorithm's ability to recognize these situations. However, for now Table 3 and Figure 21 provide preliminary support for utilizing the Euclidean distance method, as opposed to the Vowel-Space-Overlap method, during experimental data collections to best determine an appropriate set of time-independent synthesis parameters.

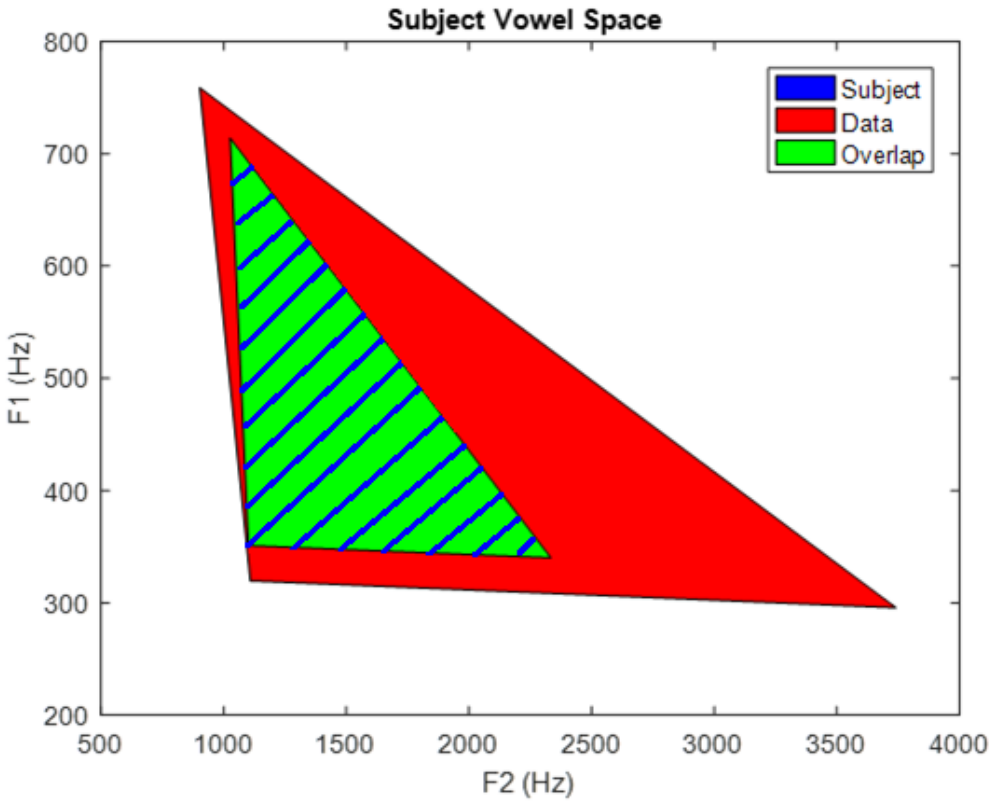

Figure 21: Vowel-Space-Overlap Method with 100% Overlap between Subject and Synthesized Formant Values

3.4 Verification of Synthesis Parameter Match to Subject's Acoustic Characteristics with Real Subject Data

3.4.1 Significance of Improved Methods

One of the key pieces of the RASS system is the VTDemo software that models the vocal tract characteristics and synthesizes speech. The current method of determining the best match between synthesis parameters and a subject's acoustic characteristics in VTDemo is through the use of a scaling factor. This scaling factor shrinks or expands the overall size of the vocal tract model to a fixed value that reflects the size of the subject's vocal tract. While a look-up table and Vowel-Space-Overlap method are used to determine the best scaling factor value, the scope of the matching method is limited by a small sample of formant values. The act of expanding the look-up table over a range of different fundamental frequencies and adding the LH parameter increases the likelihood of finding a more similar match between subject and synthesized formant values. The closer the match between formant values, the more likely that the synthesized speech will sound analogous to the subject's speech. Although in theory the idea of expanding the database and increasing the number of parameters is a logical step in improving the accuracy of a matching algorithm, it is important to verify the impact through quantitative data.

3.4.2 Verification Method

3.4.2.1 Evaluation of an Expanded Set of Vowels in MATLAB

In order to evaluate the impact of the expanded time-independent parameter-determination methods mentioned in Section 3.3.1, synthesized speech generated from subjects' kinematic data was analyzed so that a comparison could be drawn between the speech segments before and after the LH and SF parameters were changed from their

default values. This was accomplished by expanding the number of vowels being considered from the three that were used in determining the synthesis parameters to a set of six and measuring whether the formants of the additional vowels matched between the synthesized and actual subject measurements. To do this, the first and second formants of the synthesized vowels were measured and plotted on F1 vs F2 graphs along with subject formant values, creating vowel spaces containing subject and synthesized formant values.

Six vowels were analyzed to determine which method, Sum-Euclidean-Distance or Vowel-Space-Overlap, best matched time-independent synthesis parameters to a subject's acoustic characteristics. These vowels were "/i/," "/o/," "/u/," "/e/," "/a/," and "/ɛ/,"and they were labeled "111," "121," "131," "141," "151," and "161," respectively, to maintain consistent notation with current research in Marquette's Speech and Swallowing Lab. The first step to generating synthesized speech for each vowel was to examine the kinematic data files generated by the NDI Wave system. These data files contained the time stamps and the sensor positions tracked during articulation.

The next step to synthesizing the vowels was to map the kinematic data onto synthesis parameters using the quantile method previously discussed in Section 2.4. The significance of the mapping is that the synthesis parameters are direct inputs to the VTDemo software. Once the vowel files were synthesized in VTDemo using pre-determined SF and LH parameters, the saved audio was tracked by a third-party software, TF32, using a low-order LPC analysis to determine the first and second formant values of each vowel [18]. The formants were tracked in the middle of each vowel segment to avoid the non-uniformities often seen at the beginning and end of articulation.

3.4.2.2 Analysis of Subject Vowel Data

When the kinematic data files for the six vowels were created, all six vowels were recorded in one audio file. In order to separate the vowels, label files containing the start and stop times of each vowel were manually generated by the Marquette Speech and Swallowing Lab based on visual inspection of the first and last glottal pulses associated with the particular vowels. Knowing these vowel segment times, each vowel's first and second formant values were extracted using the TF32 software previously mentioned.

The significance of obtaining both the subject's original real and synthesized formant values for the six vowels is that the accuracy of the synthesizer can be observed by comparing the two sets of formants before any SF or LH parameter is modified (original baseline). As one of the subjects studied during a RASS experiment performed in Marquette's Speech and Swallowing Lab, Subject 27's vowel space with both real and synthesized formant values was examined and can be seen in Figure 22. The number labels next to each point on the plot correspond to specific vowels as defined in Section 3.4.2.1. The figure below reveals that the synthesized vowels (in red) are not located identically in the vowel space to the subject vowels (blue), revealing there is room for improvement.

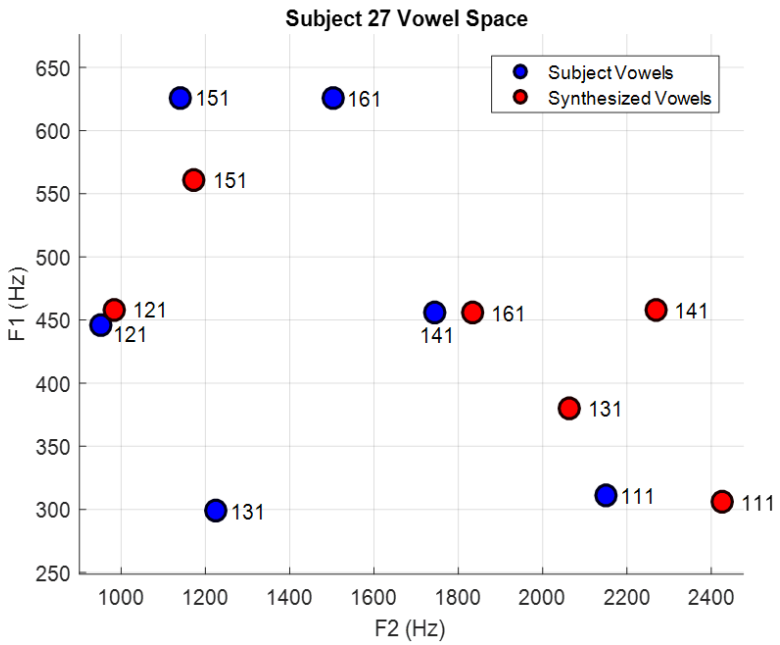

Figure 22: Original Baseline Synthesis Parameter Vowel Space Plot for Six Vowels

3.5 Results of Time-Independent Parameter Matching

The results of the synthesis-parameter-to-subject matching methods previously discussed in Section 3.4.2.3 are best analyzed by plotting the formant values in the vowel space similar to Figure 22. When the synthesized formant values are in the same plot as the subject's real formant values for vowels, comparisons can be drawn between the methods. One specific metric for analyzing the synthesized and real formant values is the Euclidean distance formula (Eq. 3.1). A smaller Euclidean distance between each synthesized vowel and respective real subject vowel signifies a higher degree of similarity between the synthesized speech and the subject's real speech, which ultimately indicates a more accurate representation of the subject's acoustic characteristics.

In order to study whether the methods for time-independent parameter determination effectively increase subject matching objectives, six variations of formant synthesis were performed. The first of these is the original baseline (OB) as previously

described in Section 3.4.2.2., where the vowel formants were synthesized when the scaling factor was set to 1 and the laryngeal height parameter to 0 (default VTDemo settings). The second case is the current baseline (CB), which is the approach Marquette's Speech and Swallowing Lab currently uses to modify the SF parameter to achieve subject-matching. This involves using the Vowel-Space-Overlap method to determine a scaling factor for the synthesized vocal tract, with the LH parameter at the default setting of 0.

The third and fourth cases of formant synthesis utilize the Sum-Euclidean-Distance (SED) and Vowel-Space-Overlap (VSO) methods, respectively, to assign subject-specific SF and LH parameters to the subject's vocal tract model as described in Section 3.3. The fifth and sixth experimental cases are identical to the third and fourth, except that the subject's average F0 of three vowels (111, 131, and 151) is used as an input to select the SF and LH values in addition to the corner vowel formants. These F0-modified SED and VSO methods only consider synthesized corner vowel formant values in the look-up table that were synthesized at an F0 differing by a maximum of 1.7 Hz from the subject's average F0.

In total, nine subjects were studied in this thesis, with subjects' kinematic data and original audio files provided by Marquette's Speech and Swallowing Lab. The vowel space plots containing the subject's real and synthesized vowels for all six method variations for six vowels can be seen in the figures below (Figures 23 to 31). Due to the quality of some audio files and synthesis capabilities of the RASS system, some vowels were not able to be correctly synthesized for the nine subjects. More specifically, the quality of the original kinematic data in combination with the mapping algorithm did not

always result in an intelligible vowel sound once synthesized. However, all six vowels were synthesized at least once across the subjects.

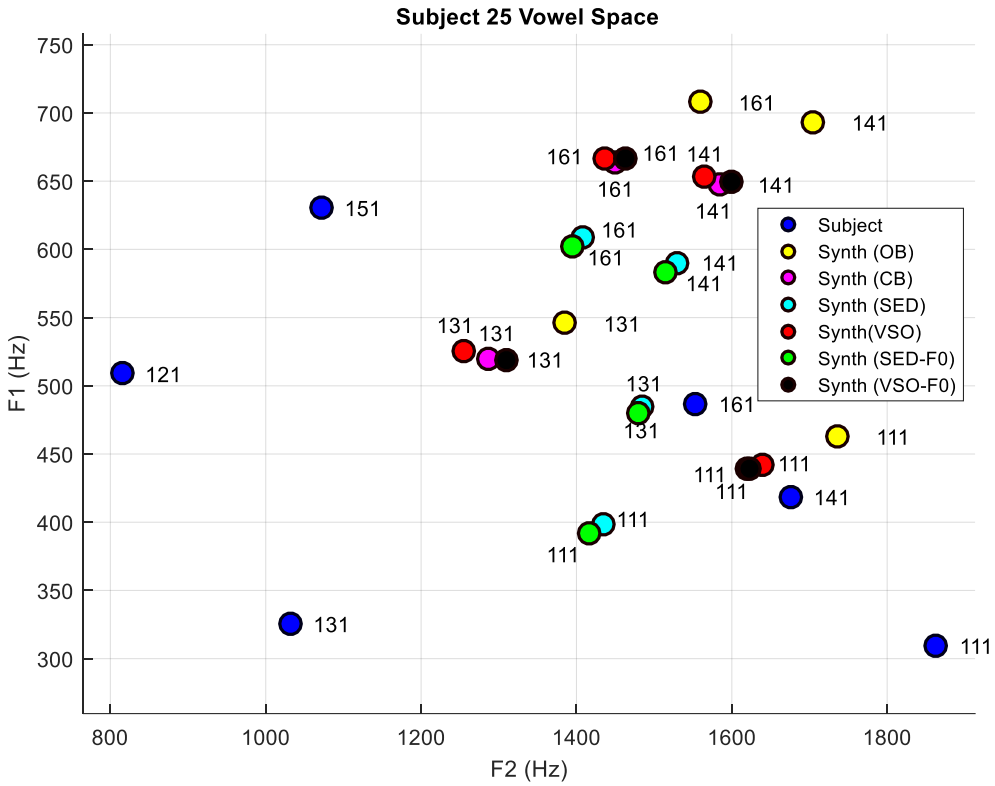

Figure 23: Vowel Space for Subject 25

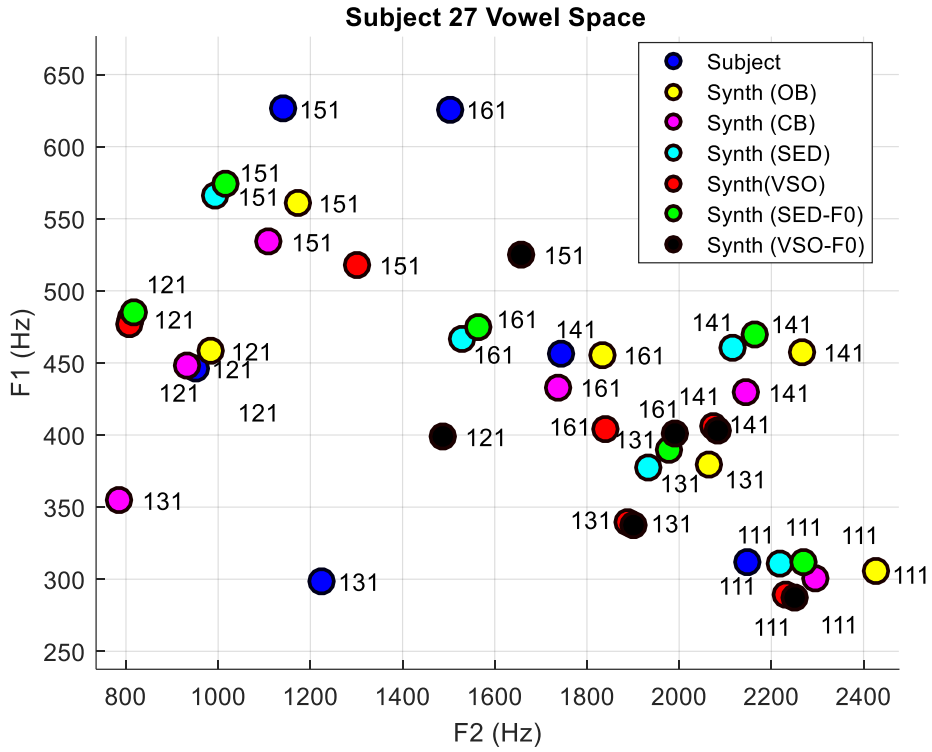

Figure 24: Vowel Space for Subject 27

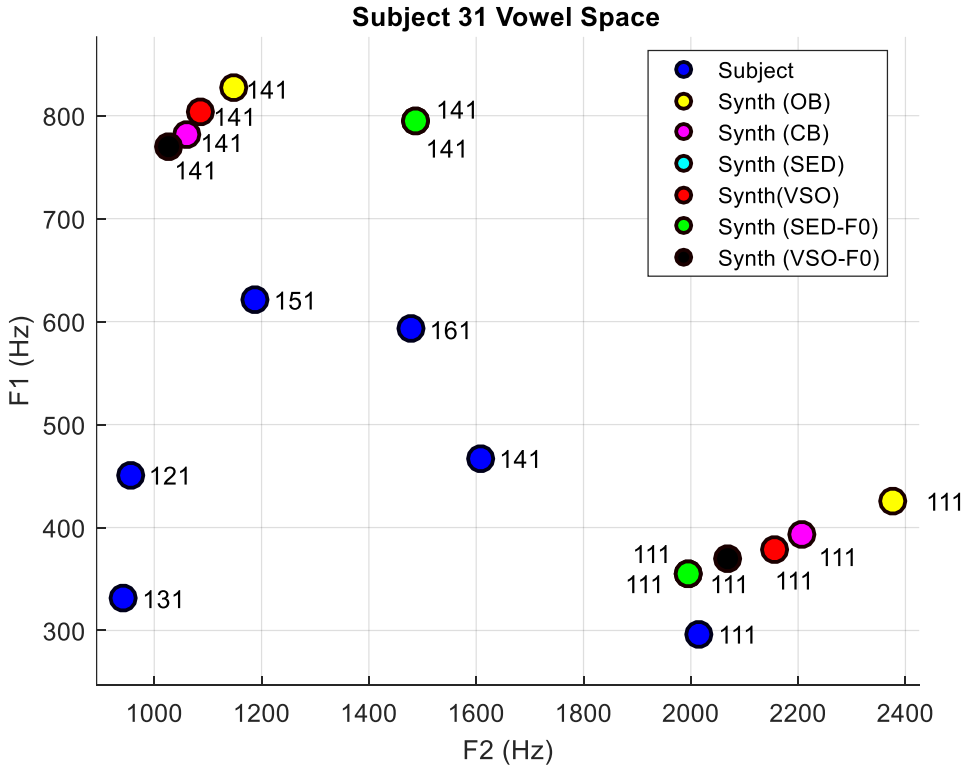

Figure 25: Vowel Space for Subject 31

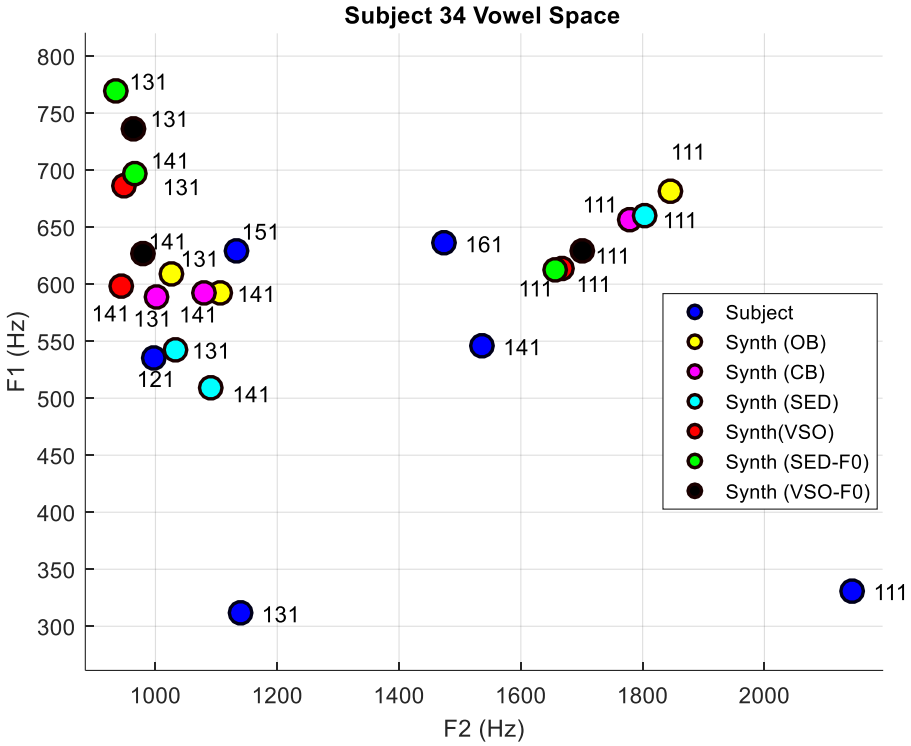

Figure 26: Vowel Space for Subject 34

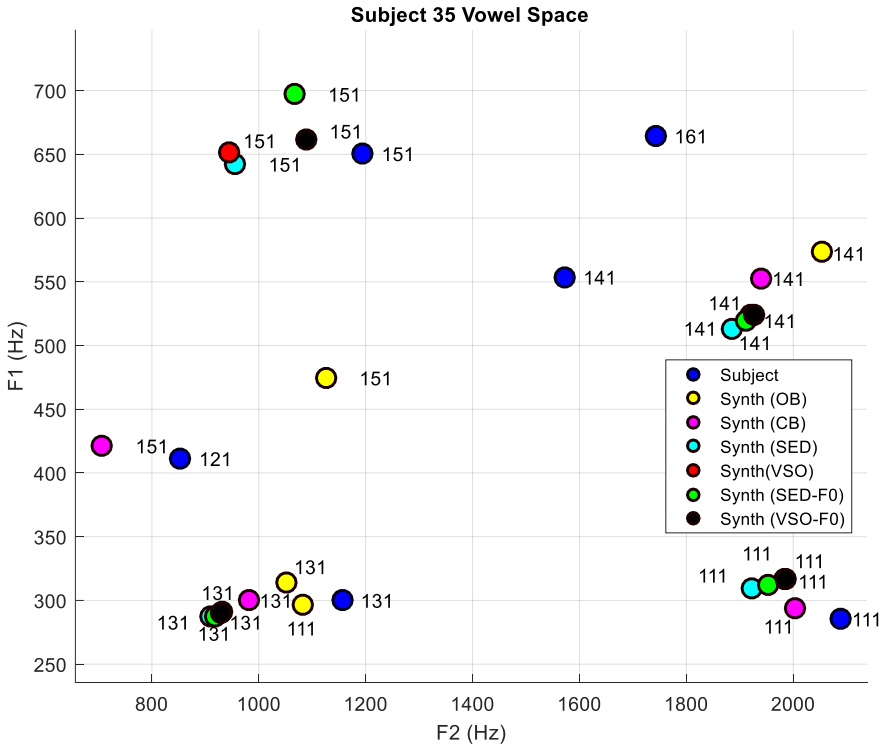

Figure 27: Vowel Space for Subject 35

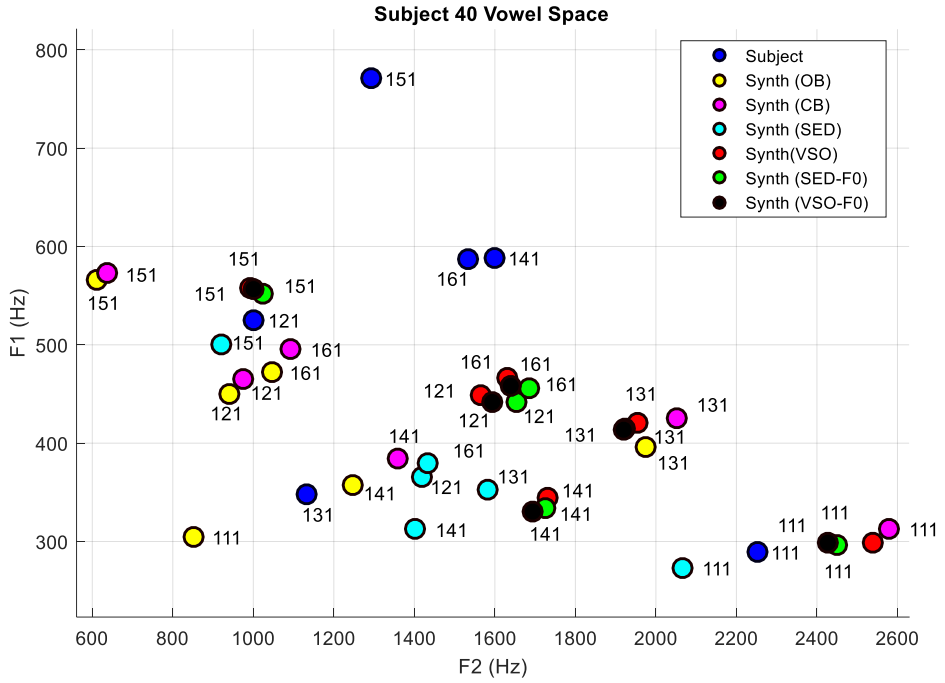

Figure 28: Vowel Space for Subject 40

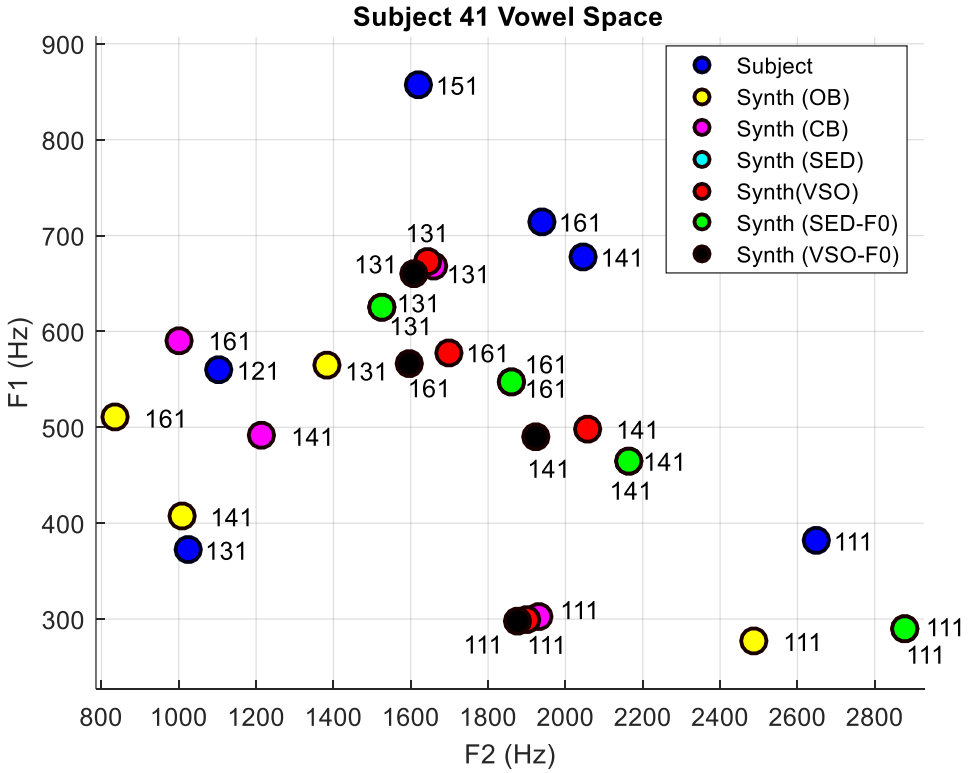

Figure 29: Vowel Space for Subject 41

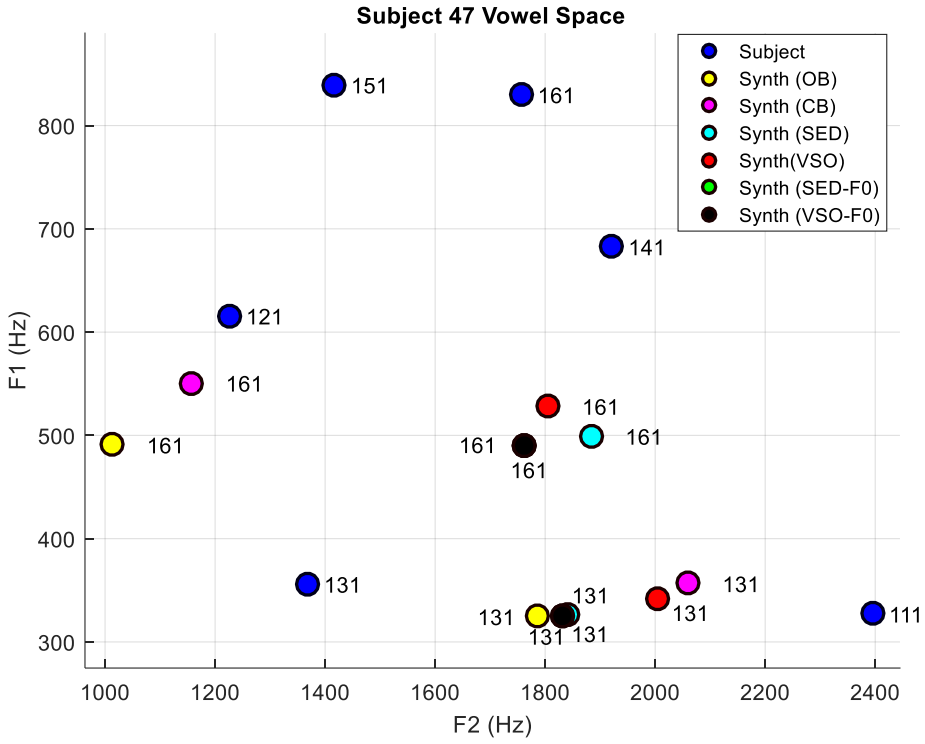

Figure 30: Vowel Space for Subject 47

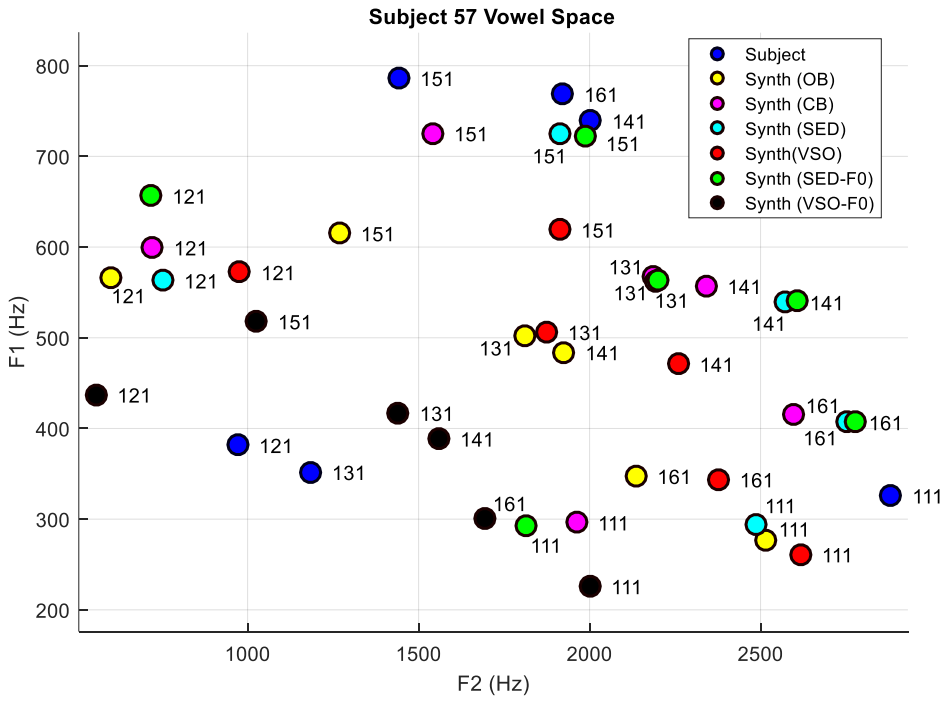

Figure 31: Vowel Space for Subject 57

From the vowel space plots alone, it can be observed that for each of the subjects there was a significant difference between the synthesized and real formants. However, since the focus of this thesis is improving the subject-matching capabilities of RASS, an important aspect to analyze is whether the newly introduced methods provided a better match than the original baseline. The Euclidean distance metric, as previously described, offers clarification and a technique by which to compare the different methods on display in the figures above. The next nine tables contain those Euclidean distances for each of the six method variations displayed in the nine subjects' vowel plots (Tables 3 through 11). The highlighted rows signify the method that has the minimum total Euclidean distance summed across the six vowels.

Table 3: Subject 25's Euclidean Distances to Analyze Parameter-Determination Methods

Subject 25 - Euclidean Distances Between Real and Synthesized Formants for Six Vowels							
	111 (Hz)	**121 (Hz)**	**131 (Hz)**	**141 (Hz)**	**151 (Hz)**	**161 (Hz)**	**Sum Distance (Hz)**
Original Baseline	199.1096	n/a	416.8593	276.0919	n/a	221.0651	1113.1259
Current Baseline	275.9640	n/a	320.3525	246.6878	n/a	204.6244	1047.6287
SED	437.5486	n/a	480.8448	224.5437	n/a	189.4747	1332.4118
VSO	259.5271	n/a	299.7339	259.4555	n/a	214.2529	1032.9694
SED with F0	454.3295	n/a	474.7374	230.4622	n/a	196.4300	1355.9591
VSO with F0	273.8943	n/a	339.6574	243.0368	n/a	201.2880	1057.8765

Table 4: Subject 27's Euclidean Distances to Analyze Parameter-Determination Methods

Subject 27 - Euclidean Distances Between Real and Synthesized Formants for Six Vowels							
	111 (Hz)	**121 (Hz)**	**131 (Hz)**	**141 (Hz)**	**151 (Hz)**	**161 (Hz)**	**Sum Distance (Hz)**
Original Baseline	277.3319	33.9335	842.1216	523.3650	72.4934	372.3899	2121.6353
Current Baseline	145.4166	18.4025	443.6625	400.0817	97.6079	302.0285	1407.1997
SED	68.0911	144.8837	712.8958	370.3843	160.4946	161.4591	1618.2086
VSO	85.7845	148.1637	664.8738	333.9385	192.2379	404.0083	1829.0067
SED with F0	120.9108	139.8621	759.5183	420.8603	135.5035	162.8717	1739.5267
VSO with F0	102.6205	536.8874	675.2077	342.5742	525.0780	537.0931	2719.4609

Table 5: Subject 31's Euclidean Distances to Analyze Parameter-Determination Methods

Subject 31 - Euclidean Distances Between Real and Synthesized Formants for Six Vowels							
	111 (Hz)	**121 (Hz)**	**131 (Hz)**	**141 (Hz)**	**151 (Hz)**	**161 (Hz)**	**Sum Distance (Hz)**
Original Baseline	381.9320	n/a	n/a	584.4396	n/a	n/a	966.3716
Current Baseline	213.4841	n/a	n/a	631.6085	n/a	n/a	845.0926
SED	61.3906	n/a	n/a	348.7498	n/a	n/a	410.1404
VSO	161.5203	n/a	n/a	621.5709	n/a	n/a	783.0912
SED with F0	61.3906	n/a	n/a	348.7498	n/a	n/a	410.1404
VSO with F0	89.9654	n/a	n/a	655.1180	n/a	n/a	745.0834

Table 6: Subject 34's Euclidean Distances to Analyze Parameter-Determination Methods

Subject 34 - Euclidean Distances Between Real and Synthesized Formants for Six Vowels							
	111 (Hz)	**121 (Hz)**	**131 (Hz)**	**141 (Hz)**	**151 (Hz)**	**161 (Hz)**	**Sum Distance (Hz)**
Original Baseline	460.0936	n/a	318.6367	431.1660	n/a	n/a	1209.8963
Current Baseline	489.0784	n/a	309.5940	459.7167	n/a	n/a	1258.3891
SED	473.8202	n/a	254.4384	447.0417	n/a	n/a	1175.3003
VSO	555.2414	n/a	420.1035	594.4764	n/a	n/a	1569.8213
SED with F0	563.7202	n/a	502.3015	589.9129	n/a	n/a	1655.9346
VSO with F0	535.1886	n/a	459.3970	562.3827	n/a	n/a	1556.9683

Table 7: Subject 35's Euclidean Distances to Analyze Parameter-Determination Methods

Subject 35 - Euclidean Distances Between Real and Synthesized Formants for Six Vowels							
	111 (Hz)	**121 (Hz)**	**131 (Hz)**	**141 (Hz)**	**151 (Hz)**	**161 (Hz)**	**Sum Distance (Hz)**
Original Baseline	1006.7894	n/a	104.9686	480.8520	188.9909	n/a	1781.6009
Current Baseline	85.8367	n/a	173.1819	367.4562	539.1753	n/a	1165.6501
SED	166.4161	n/a	247.3184	316.4266	239.1431	n/a	969.3042
VSO	108.6526	n/a	225.3803	349.8498	249.2748	n/a	933.1575
SED with F0	137.3000	n/a	236.7872	341.3980	135.8043	n/a	851.2895
VSO with F0	106.5061	n/a	225.6885	356.2123	105.8746	n/a	794.2815

Table 8: Subject 40's Euclidean Distances to Analyze Parameter-Determination Methods

Subject 40 - Euclidean Distances Between Real and Synthesized Formants for Six Vowels							
	111 (Hz)	**121 (Hz)**	**131 (Hz)**	**141 (Hz)**	**151 (Hz)**	**161 (Hz)**	**Sum Distance (Hz)**
Original Baseline	1398.2662	97.2732	843.6758	419.2854	710.8205	500.3104	3969.6315
Current Baseline	328.8571	65.7123	923.5048	314.4016	685.0740	449.2595	2766.8093
SED	186.4230	447.9755	451.8490	338.4833	459.8237	229.7959	2114.3504
VSO	286.7137	568.2943	827.7364	278.2639	367.0049	155.5668	2483.58
SED with F0	198.9545	658.1100	793.8402	284.5513	346.7599	200.7523	2482.9682
VSO with F0	177.2287	598.7663	791.5745	274.2259	361.6282	168.1855	2371.6091

Table 9: Subject 41's Euclidean Distances to Analyze Parameter-Determination Methods

Subject 41 - Euclidean Distances Between Real and Synthesized Formants for Six Vowels							
	111 (Hz)	**121 (Hz)**	**131 (Hz)**	**141 (Hz)**	**151 (Hz)**	**161 (Hz)**	**Sum Distance (Hz)**
Original Baseline	192.0380	n/a	408.2432	1074.0667	n/a	1121.0905	2795.4384
Current Baseline	722.4832	n/a	700.2984	851.4058	n/a	944.3224	3218.5098
SED	246.0164	n/a	561.7460	243.9971	n/a	184.6266	1236.3861
VSO	752.8020	n/a	691.2880	179.8340	n/a	274.4551	1898.3791
SED with F0	246.0164	n/a	561.7460	243.9971	n/a	184.6266	1236.3861
VSO with F0	779.2287	n/a	651.0736	224.1030	n/a	372.6841	2027.0894

Table 10: Subject 47's Euclidean Distances to Analyze Parameter-Determination Methods

Subject 47 - Euclidean Distances Between Real and Synthesized Formants for Six Vowels							
	111 (Hz)	**121 (Hz)**	**131 (Hz)**	**141 (Hz)**	**151 (Hz)**	**161 (Hz)**	**Sum Distance (Hz)**
Original Baseline	n/a	n/a	418.3591	n/a	n/a	816.7200	1235.0791
Current Baseline	n/a	n/a	691.6366	n/a	n/a	661.3985	1353.0351
SED	n/a	n/a	473.9219	n/a	n/a	354.1756	828.0975
VSO	n/a	n/a	635.7140	n/a	n/a	305.5358	941.2498
SED with F0	n/a	n/a	463.6860	n/a	n/a	340.1189	803.8049
VSO with F0	n/a	n/a	463.6860	n/a	n/a	340.1189	803.8049

Table 11: Subject 57's Euclidean Distances to Analyze Parameter-Determination Methods

Subject 57 - Euclidean Distances Between Real and Synthesized Formants for Six Vowels							
	111 (Hz)	**121 (Hz)**	**131 (Hz)**	**141 (Hz)**	**151 (Hz)**	**161 (Hz)**	**Sum Distance (Hz)**
Original Baseline	369.0270	415.3806	642.9852	266.9164	243.7647	472.3876	2410.4615
Current Baseline	919.2990	331.0226	1023.3478	383.2517	116.2365	763.1152	3536.2728
SED	396.8378	284.6160	1028.9678	601.6128	473.6166	907.3412	3692.9922
VSO	272.2155	190.7365	706.8785	370.4986	499.3624	622.5490	2662.2405
SED with F0	1067.5102	374.0407	1039.2288	635.5357	550.1580	929.1984	4595.6718
VSO with F0	885.8442	418.2623	264.8727	565.9305	494.9249	521.1543	3150.9889

The Euclidean distance between synthesized and real vowels is a significant metric because it takes both the first and second formants into account instead of comparing individual raw formant values to each other. Methods with the smallest sum of Euclidean distances across the six vowels for each subject reveals the highest degree of similarity between synthesis parameters and the subject's acoustic characteristics. A more concise summary can be seen below in Tables 12 and 13. The red text in Table 13

signifies the method that best minimized the Euclidean distance on a vowel-by-vowel basis.

Table 12: Subjects' Methods that Minimize Euclidean Distance

Subjects and Respective Methods That Minimize the Euclidean Distancess Between Real and Synthesized Formants					
Original Baseline	**Current Baseline**	**SED**	**VSO**	**SED with F0**	**VSO with F0**
57	27	31	25	31	35
		34		41	47
		40		47	
		41			

Table 13: Average of Nine Subjects' Euclidean Distances for Each Method

Averages of All Nine Subjects - Euclidean Distances Between Real and Synthesized Formants for Six Vowels							
	111 (Hz)	**121 (Hz)**	**131 (Hz)**	**141 (Hz)**	**151 (Hz)**	**161 (Hz)**	**Sum Distance (Hz)**
Original Baseline	535.5735	182.1958	499.4812	507.0229	304.0174	583.9939	2612.284583
Current Baseline	397.5524	138.3791	573.1973	456.8263	359.5234	554.1248	2479.603258
SED	254.5680	292.4917	526.4978	361.4049	333.2695	337.8122	2106.044067
VSO	310.3071	302.3982	558.9636	373.4860	326.9700	329.3947	2201.519454
SED with F0	356.2665	390.6709	603.9807	386.9334	292.0564	335.6663	2365.574288
VSO with F0	368.8096	517.9720	483.8947	402.9479	371.8764	356.7540	2502.254571

Based on the results of the experiments in this chapter, the Sum-Euclidean-Distance (SED) method of time-independent parameter determination best minimizes the Euclidean distance between subjects' real and synthesized vowels. Since this method is designed to minimize the sum of the Euclidean distances for synthesis parameter selection, it is reasonable that it would generally produce the best results when Euclidean distance is the metric for evaluating the above vowel space plots. One may note, however, that the SED method was not the best method in every case, and this is likely due to the fact that the plots in this section compare real subject formants versus synthesized subject formants, and the SED method compared real subject formants versus artificial subject formants (preset vowel settings previously mentioned in Section 3.2.2). Since the 121, 141, and 161 vowels were not included in the SED method, the fact that they commonly displayed a minimized Euclidean distance when the SED method was

selected aids in validating the SED method of synthesis parameter determination. Some subjects may be listed in more than one column, which indicates that the same SF and LH was determined by those different methods. For example, Subject 31 appears in both the "SED" column and under "SED with F0."

A related observation is that the inclusion of a subjects' average F0 to aid in the selection of SF and LH parameters did not play a significant role. There are five subjects listed in the combined columns for SED and VSO that didn't use the F0 input and five in the columns where F0 was utilized. The time-independent (fixed) parameters under investigation in this thesis that were used with the methods in Table 12 can be seen below in Table 14 along with each subject's average vowel F0.

Table 14: Appropriate Subject-Specific Synthesis Parameters for Subject Matching

Subject Number	Scaling Factor	Laryngeal Height	Real Average F0 (Hz)
25	1.06	-0.60	107.7158
27	1.06	0	146.2623
31	1.26	2.90	117.7829
34	0.98	-2.80	147.5539
35	1.18	2.40	135.6612
40	1.10	2.70	110.2125
41	0.94	2.90	173.2314
47	1.04	2.60	281.2906
57	1.00	0	213.9435

3.6 Conclusion Based on Time-Independent Parameter Determination Results

This chapter showed that the SED method most frequently provides the best formant match between synthesis parameters and subjects. Another conclusion that can be drawn is that the SED and VSO methods, both with and without the additional F0 input, provided an overall better representation of the subject's acoustic characteristics than the current and original baselines. Specifically looking at an outlier in Subject 57's

results, it is likely that the quality of the subject's original vowel audio recordings and kinematic data caused the original baseline method to minimize the sum of the Euclidean distance across six vowels. Similar outliers can be seen on a smaller scale such as Subject 27's 131 vowel, where the real vowel is located far from the synthesized 131 vowels in the formant space.

RASS encompasses the capture of kinematic data gathered from articulator movements, maps those data to synthesis parameters, and ultimately sends the synthesized audio back to the subject. The time-independent parameter-determination methods studied in this chapter, Sum-Euclidean-Distance and Vowel-Space-Overlap, play a significant role in improving the subject-matching capabilities of the system by introducing the use of subject-specific SF and LH combinations. For the purposes of rehabilitation and achieving learning outcomes, the best selected subject-specific synthesis parameters promote increased involuntary learning through sensorimotor adaptation.

CHAPTER 4: TIME-DEPENDENT PARAMETER MATCHING

4.1 Background

4.1.1 Introduction to Real-Time Parameter Tracking

While time-independent parameters have been shown in Chapter 3 to affect the outcome of synthesized speech, the application of time-dependent parameters can be used in a similar way to improve the matching of synthesis parameters to a subject's acoustic characteristics. In order to study the effect, one must identify the unused time-dependent parameters employed by the VTDemo software in RASS: FX (F0), GA (glottal aperture), and NS (nasality). These three parameters, described in Table 1, are currently set at neutral values when the RASS experiments are run. The glottal aperture and nasality parameters are set to 0, and the F0 parameter is set to 140 Hz, 230 Hz, or 270 Hz for males, females, or children, respectively. For the purpose of this thesis, FX was the only time-dependent parameter investigated, however, that does not mean that the glottal aperture and nasality have negligible effects on the quality of synthesized speech.

It is important to note that setting FX to match individual speakers in real-time is the main goal of this chapter. Doing so will not necessarily align the synthesized formant values more closely with the subject's formants as studied in Chapter 3. Therefore, while the following experiment seeks to improve the similarity between subject and synthesized speech using real-time F0-tracking, the best metric for determining whether adjusting the synthesized F0 increases involuntary learning outcomes is dependent on future subjective perceptual experiments.

One way to control the VTDemo F0 variables in real-time is to use the subject's audio and run an F0-tracking algorithm to extract the appropriate FX parameter at each

speech frame while simultaneously synthesizing the subject's speech. However, the subject's audio signal recorded during experiments does not currently pass through the NDI Wave System into RASS. In order to directly control the FX parameter, one would need to use a real-time stream from the electroglottograph (EGG) system attached to the subject during experiments. Analyzing the subject's EGG waveform would allow the F0 to be calculated, converted into FX parameters, and synthesized with the other VTDemo parameters. However, the current RASS configuration only contains a two-channel audio card, and both channels are already occupied by the EMA data and SMPTE timecode. Since a third channel is required to facilitate the F0-tracking algorithm in real-time, a multi-channel audio card would need to be installed, and that is out of the scope of this thesis. Regardless of current RASS equipment, real-time parameter matching is still significant because it will provide researchers with a sense of the benefit of voice source control. Therefore, an offline demonstration of F0-tracking was performed in MATLAB to show the feasibility of controlling the FX parameter in VTDemo and the effects of matching the synthesis parameter to a subject's acoustic characteristics.

4.1.2 Electroglottograph Background and Use in F0-Tracking

The demonstration mentioned was designed to show the feasibility of changing the FX parameter in real-time based on the fundamental frequencies gathered from a subject's EGG signal. As previously mentioned, the EGG system is not currently compatible with real-time F0-tracking in the RASS system. However, a better understanding of the EGG system and how to use the EGG signal for real-time F0-tracking provides a context for the MATLAB demonstration. The electroglottograph (EGG) can be utilized in order to non-invasively estimate the degree of abduction and

adduction of the vocal folds during voiced speech. These values can be approximated from the variations in vocal fold contact area (VFCA) that occur as the vocal folds vibrate. Abduction of vocal folds results in a smaller VFCA and shorter contact period. As long as the vocal folds remain in contact, the degree of abduction can be estimated [19].

The EGG operates through the use of transverse electrical conductance (TEC). Two electrodes are positioned on either side of the neck at the level of the larynx. A small AC current operating at several megahertz is passed through the neck from one electrode and received by the other. During voiced speech, the vocal folds come together, yielding an increase in TEC that is then recorded by the EGG system and interpreted by the user as an instant of contact. A larger increase in TEC denotes a larger contact area, although the increase in TEC is usually only on the order of magnitude of 1% of the total conductance. The measured conductance may vary according to the subject's neck anatomy, including the position of the glottis, structure of thyroid cartilage, and the amount of muscular, glandular, and fatty tissue around the larynx. Additionally, the electrode configuration and placement will increase variation, with another source of error being the depth to which the electrodes are pressed into the subcutaneous fatty tissue of the neck [19]. Figure 32 illustrates the correct orientation of the electrodes on the neck and indicates the method of collecting EGG data.

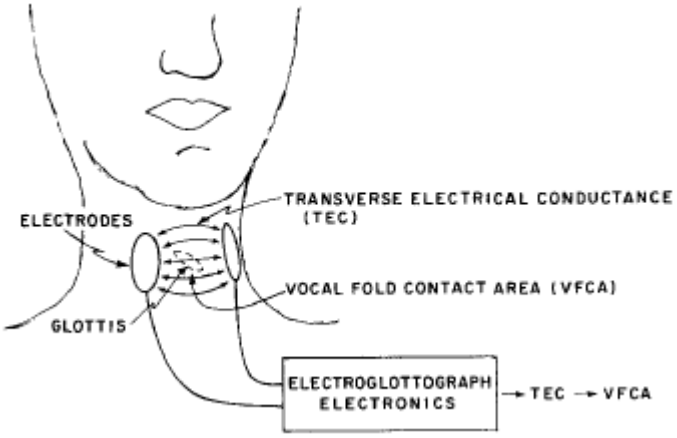

Figure 32: Sketch of the Correct Electrode Placement and Data Collection [19]

The resulting EGG signal yields information about the change in VFCA with respect to time. Figure 33 represents an idealized EGG waveform with the vocal fold events labeled to correspond to the sketches of the vocal fold motion below.

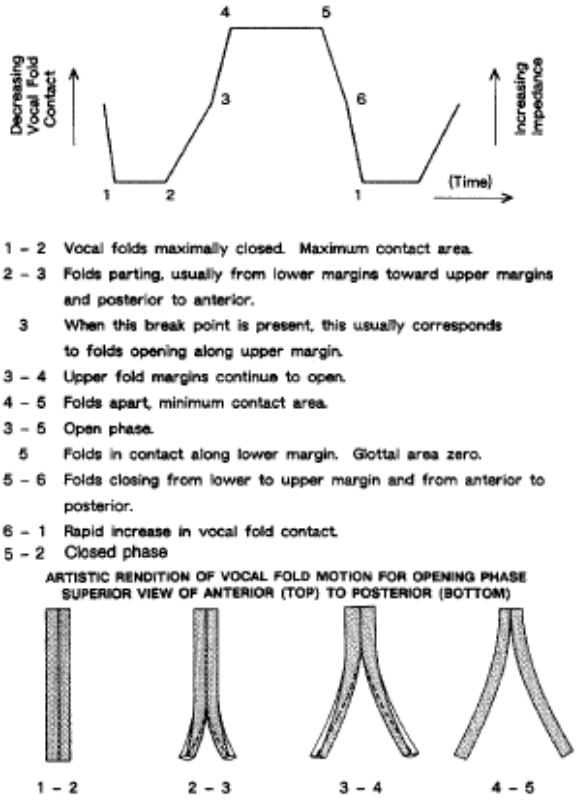

Figure 33: Ideal EGG Waveform with Corresponding Vocal Fold Events [20]

The EGG waveform can be roughly characterized by a mathematical model seen in Equation 4.1.1:

$$EGG(t) = k / [A(t) + C] \tag{4.1.1}$$

in which *t* represents time, *k* represents a scaling constant, *A(t)* represents the vocal fold contact area, and *C* is a constant that is proportional to the shunt impedance at *A(t)* = 0 [20].

One application for the data gathered by the EGG is F0 detection, which can be performed utilizing the differentiated electroglottograph signal, also referred to as DEGG, as seen in Figure 34.

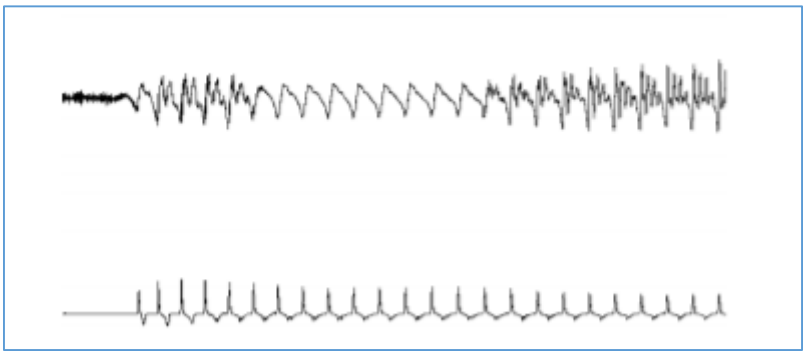

Figure 34: EGG Waveform (top) and DEGG Waveform (bottom) [21]

Each positive peak of the DEGG signal indicates that a glottal closure instant (GCI) has taken place. The time difference between GCIs is referred to as the F0 period. The F0 can then be calculated by taking the inverse of the F0 period. In order to correctly identify periods of voiced speech, a threshold can be applied to the DEGG signal. It is suggested that this threshold be found by examining the DEGG signal during moments in which the subject was silent [21].

4.2 F0-Tracking Demonstration with Acoustic Signal and FX Parameter

As the previous section details, a subject's F0 can be tracked in real-time based on the characteristics of the subject's EGG signal. The intended application of EGG-based F0-tracking is to convert the calculated F0 values into FX parameter values for subject matching applications in VTDemo. A fluid FX parameter that accurately matches VTDemo's synthesized speech to the subject's F0 in real time likely allows for a better match between synthesis parameters and a subject's acoustic characteristics, which is the overall goal of this thesis in support of involuntary learning outcomes.

In order to model the real-time F0-tracking and use of the FX parameter, a demonstration was created using a third-party MATLAB F0-tracker and acoustic recordings of the previously studied subjects in Chapter 3 [22]. The F0-tracking algorithm estimates the F0 value every speech frame within the range of 75-500 Hz. Additionally, the spectrum is uniformly sampled every $1/20^{th}$ of ERB (equivalent rectangular bandwidth) and a Hann window of 50% overlap is used. For fine tuning the F0, a parabolic interpolation algorithm is used and low strength F0 estimates are treated as undefined. The F0 trace plot seen in Figure 35 is a generic representation of a subject's F0 values over time for a vowel before they are converted into FX parameters.

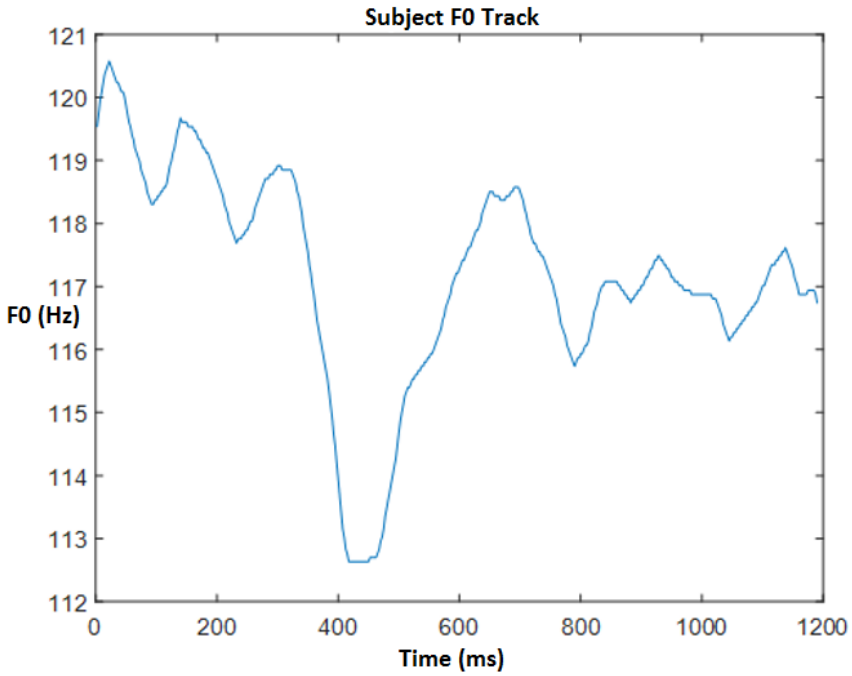

Figure 35: F0 Estimation from a Subject's Acoustic Signal Using MATLAB

The F0 estimation algorithm sends the F0 of each frame to an array where each entry is assigned to one of the FX parameter integer intervals between -4.0 and 10.0 based on the F0 values in Hertz. Once each frame's F0 value is sorted into an integer interval, a linear interpolation is performed between FX integer values that are spaced 17 Hz apart as mentioned in Chapter 2. These newly calculated FX parameter values can then be loaded into VTDemo for synthesis. It is important to note that even though this demonstration is performed outside of RASS, the configuration is still sufficient to study the effects of time-dependent synthesis parameters. The concept of tracking F0 and determining the FX parameters in the VTDemo software for speech synthesis can be employed in future RASS system configurations that use an EGG signal.

4.3 Verification of the Time-Dependent FX Parameter with Real Subjects

Implementation of the time-dependent FX parameter for real-time F0-tracking in VTDemo is difficult to evaluate without subjective perceptual experiments. In an experimental setting, subjects would gauge whether or not the synthesized speech played back to them sounds more like their natural voice than without the F0-track. Since RASS is not configured to accept EGG inputs, this experiment is not currently possible. However, to verify that the algorithm to convert a subject's F0 to FX parameters is functional, the F0-track was performed on recorded audio files for the nine subjects previously studied. For this experiment, the phrase "I owe you a yo-yo" was used for synthesis due to its composition of voiced sounds. The F0 values gathered from the subject's real audio files for this phrase were converted to FX parameter values at the appropriate time segments, aligning with variations in the subject's F0. Using the time-independent synthesis parameters in Table 12 along with newly recorded FX parameter values, the subject's speech was then synthesized in VTDemo and exported to audio files for purposes of perceptive evaluation. This will allow one to compare real audio, synthesized audio without the F0-track, and synthesized audio with the F0-track.

One way to quantify the F0-tracking algorithm designed for RASS is to simultaneously plot the F0 of both the real and synthesized audio over time. Due to the nature of this experiment, demonstration rather than data collection, only three of the F0-track plots are shown below. The remaining six subjects' plots are very similar and located in Figures 39 to 44, Appendix B.

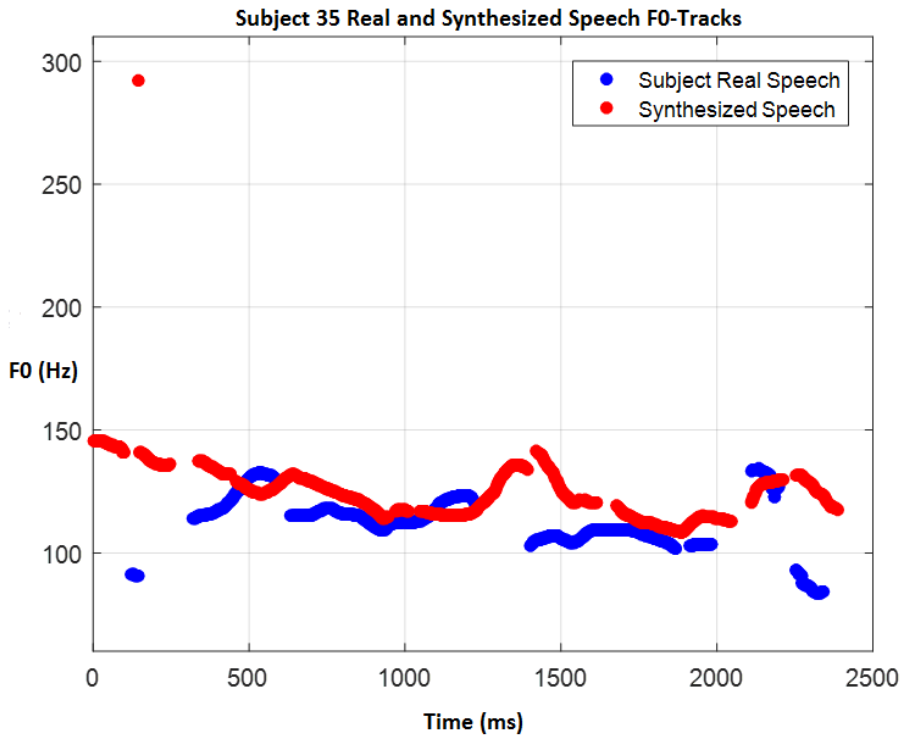

Figure 36: F0-Track of Subject 35's Real and Synthesized Speech

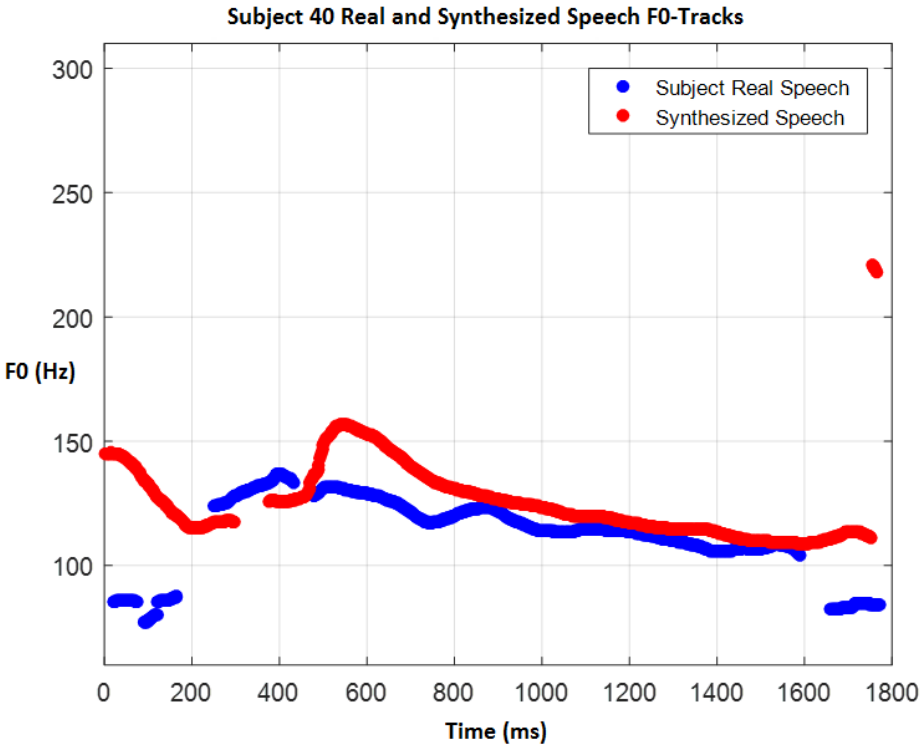

Figure 37: F0-Track of Subject 40's Real and Synthesized Speech

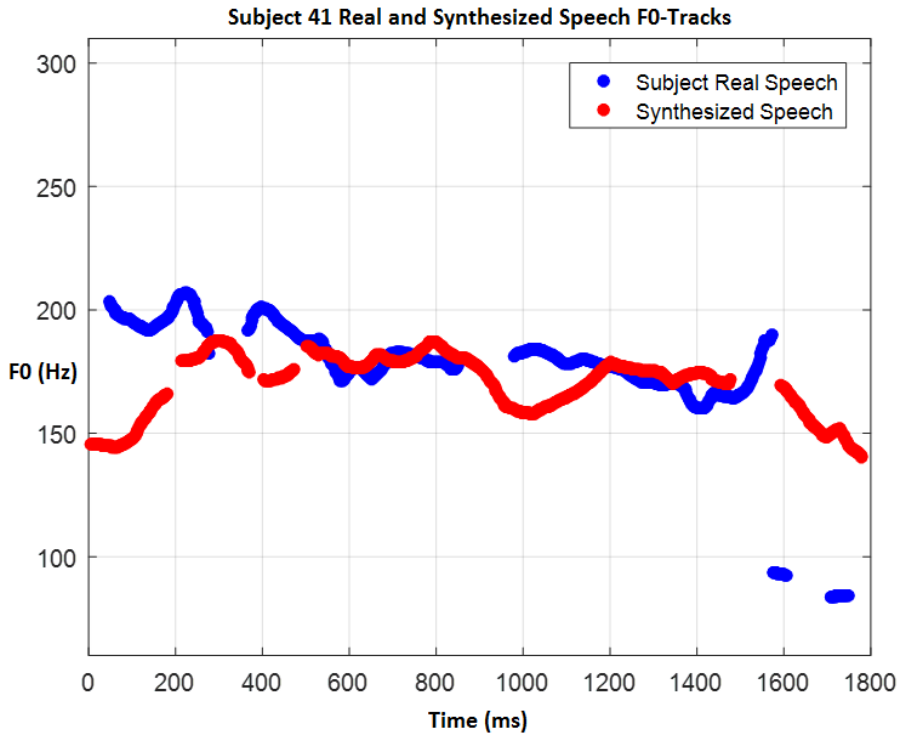

Figure 38: F0-Track of Subject 41's Real and Synthesized Speech

As can be seen in the three figures above, the synthesized F0 parameter sufficiently follows the subject's real F0. This level of tracking suggests that users should be able to perceive an increased auditory similarity between the subjects' real and synthesized speech, which ultimately opens the door to a higher degree of involuntary learning in future experiments. However, one may note that the figures containing F0-tracks do not display perfect alignment between the real and synthesized F0 values over the course of the speech segment. This inaccuracy is important to note but is likely due to the original quality of synthesized speech, which ultimately relates back to the kinematic articulatory data and mapping mentioned in Chapter 3. Another possible source of error is the third-party F0-tracker's algorithmic inaccuracies, which could play a role in generating inconsistent F0 values. This type of error could also be present when

a real EGG signal is used, depending on the quality of the EGG signal processing. While the F0-track shows generally consistent patterns between the subjects' real and synthesized F0 values over time (Figures 36 through 38), future perceptual experiments will fully determine if there are advantages to the F0-tracking method versus using an average, constant F0 for the FX parameter.

4.4 Conclusions of Time-Dependent Parameter Synthesis

The purpose of the F0-track in this thesis was to act as demonstration for applying the EGG signal to better control the FX parameter. Implementing an F0-tracking algorithm on an EGG signal would provide a cleaner representation of the FX parameter than the demonstration's algorithm because background audio noise and sound quality of the recording would not be a factor. The FX parameter would also be exclusively derived from the movements the subject's vocal folds, which provides a direct connection between the subject's acoustic characteristics and synthesis parameter. Overall, this demonstration shows that the F0-track algorithm which controls the FX parameter in real-time is potentially beneficial to increasing involuntary learning outcomes.

CHAPTER 5: CONCLUSION

5.1 Summary of Thesis Work

The work in this thesis analyzed the current configuration of the RASS system in Marquette's Speech and Swallowing Lab and provided an improved algorithm to match synthesis parameters in VTDemo to a subject's acoustic characteristics. With the overall goal of increasing involuntary learning through acoustic feedback mechanisms, the enhanced methods of determining synthesis parameters increased the potential for this to be experienced by subjects.

The determination of synthesis parameters was divided into two categories: time-independent and time-dependent. The time-independent parameters under analysis were the scaling factor and laryngeal height. These two parameters control the shape and size of the modeled vocal tract in VTDemo, which has a direct impact on speech synthesis. Altering synthesis parameters changes the formant values of the synthesized speech, which are characteristic to each individual subject. Two specific methods, Sum-Euclidean-Distance and Vowel-Space-Overlap were studied to determine the best method of time-independent parameter determination, which resulted in the Sum-Euclidean-Distance method after studying nine subjects.

The second category of synthesis parameters, time-dependent variables, focused on the ability to control the FX (F0) parameter to match the subject in real-time during speech synthesis. For this thesis, a demonstration of FX parameter determination was performed using a third-party F0-tracker on audio clips of subjects' speech. F0-track plots compared subjects' real and synthesized F0 over time for a speech segment using synthesis parameters (SF and LH) derived in Chapter 3. Since the real-time F0-tracking

produced similar results between both the subjects' real and synthesized audio, it was determined that the time-dependent use of the FX parameter would potentially be a useful tool for increasing the correspondence between the synthesized speech and subjects' acoustic characteristics. Audio files containing both the real and synthesized speech segments from the plots employing the time-varying FX parameter can also be used in future work to confirm the perceptual similarity between subjects and their synthesized audio.

5.2 Contributions to Research

This thesis provides four main contributions to the research conducted in Marquette's Speech and Swallowing lab. The first contribution is the development of a database of VTDemo formant values for three synthesized vowels (/i/, /a/, and /u/) in RASS. The database contains the first three formant values of each vowel synthesized across a range of varying LH, SF, and FX parameters. The second contribution to research is the development of four time-independent (SF and LH) parameter-determination algorithms which utilize Euclidean distance sums and overlapping vowel space techniques. After testing the parameter-determination algorithms on nine subjects, the Sum-Euclidean-Distance method was shown to have performed best most consistently. The third contribution is the vowel space plots containing synthesized and real subject vowels which prove to be an effective resource for future research. The fourth contribution is the demonstration of the real-time implementation of FX, a time-dependent F0 parameter, in the RASS system. This parameter matching allows researchers to gain a sense of the benefit of voice source control. These four

contributions together give researchers the opportunity to increase the capacity for involuntary learning in their experiments with RASS.

5.3 Future Work

There is potential for further work based on the results detailed in this thesis. The first opportunity is to expand the database of formant values to a higher resolution. Currently, the scaling factor is utilized in 0.02 increments from 0.8 to 1.3, and the LH parameter is incremented by 0.1 from -3.0 to 3.0. A larger database with more precise values could produce a more accurate match of LH and SF parameters between the synthesizer and subjects' acoustic characteristics. Another opportunity to advance this research is to implement the real-time FX parameter based on the subject's EGG signal, as previously discussed. This method would allow for the use of time-dependent parameters in RASS and could be a stepping stone to introducing the nasality (NS) and glottal aperture (GA) parameters as well. Finally, additional experiments could be performed with the generated audio files from the F0-tracking algorithm to determine the degree of increased perceptual similarity between subjects' real and synthesized speech.

BIBLIOGRAPHY

[1] "Statistics on Voice, Speech, and Language," National Institute on Deafness and Other Communication Disorders (NIDCD), 7 June 2010. [Online]. Available: https://www.nidcd.nih.gov/health/statistics/statistics-voice-speech-and-language. [Accessed 19 May 2016].

[2] Y. Yunusova, G. Weismer, J. R. Westbury and M. J. Lindstrom, "Articulatory Movements During Vowels in Speakers With Dysarthria and Healthy Controls," *Journal of Speech, Language, and Hearing Research*, vol. 51, pp. 596-611, June 2008.

[3] J. Berry, C. North, B. Meyers and M. T. Johnson, "Speech Sensorimotor Learning through a Virtual Vocal Tract," in *ICA 2013 Montreal*, Montreal, 2013.

[4] X. Zhou, "Least-Squares Mapping from Kinematic Data to Acoustic Synthesis Parameters for Rehabilitative Acoustic Learning," Marquette University, Milwaukee, 2016.

[5] X. Huang, A. Acero and H.-W. Hon, in *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*, Prentice Hall, 2001, pp. 21-27.

[6] R. Nave, "HyperPhysics," Georgia State University, 2016. [Online]. Available: http://hyperphysics.phy-astr.gsu.edu/hbase/Music/vowel2.html. [Accessed 5 October 2016].

[7] "VTDemo User Guide, Version 2.0," Milwaukee, 2012.

[8] A. Kolb, "Software Tools and Analysis Methods for the Use of Electromagnetic Articulography Data in Speech Research," Marquette University, Milwaukee, 2015.

[9] "Wave User Guide," Northern Digital Inc., Waterloo, 2010.

[10] M. Huckvale, "VTDemo - Vocal Tract Acoustics Demonstrator," University College London, London, 2015.

[11] R. Patel, K. Connaghan, D. Franco, E. Edsall, D. Forigt, L. Olsen, R. Lianna, E. Tyler and S. Russel, ""The Caterpillar": A Novel Reading Passage for Assessment of Motor Speech Disorders," *American Journal of Speech-Language Pathology*, vol. 22, pp. 1-9, 2013.

[12] S. Cai, S. S. Ghosh, F. H. Guenther and J. S. Perkell, ""Adaptive auditory feedback control of the production of the formant trajectories in the Mandarin triphthong /iau/ and its patterns of generalization," *Journal of the Acoustical Society of America*, vol. 128, pp. 2033-2048, 2010.

[13] J. Berry, C. North and M. T. Johnson, "Sensorimotor Adaptation of Speech Using Real-Time Articulatory Resynthesis," in *2014 IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, Milwaukee, 2014.

[14] J. Berry, J. Jaeger IV, M. Wiedenhoeft, B. Bernal and M. T. Johnson, "Consonant Context Effects on Vowel Sensorimotor Adaptation," in *Interspeech 2014*, Singapore, 2014.

[15] L. Goldstein, "Formant Frequencies," 18 November 2015. [Online]. Available: http://sail.usc.edu/~lgoldste/General_Phonetics/Source_Filter/SFc.html. [Accessed 1 July 2016].

[16] R. Singh, D. Gencaga and B. Raj, "Formant Manipulations in Voice Disguise by Mimicry," IEEE, Pittsburgh, 2016.

[17] Reference, [Online]. Available: https://www.reference.com/science/frequency-range-human-speech-3edae27f8c397c65. [Accessed 5 July 2016].

[18] P. Milenkovic, 26 July 2005. [Online]. Available: http://userpages.chorus.net/cspeech/. [Accessed 16 November 2016].

[19] M. Rothenberg and J. J. Mahshie, "Monitoring Vocal Fold Abduction Through Vocal Fold Contact Area," *Journal of Speech and Hearing Research*, vol. 31, pp. 338-351, 1988.

[20] D. G. Childers, D. M. Hicks, G. P. Moore, L. Eskenazi and A. L. Lalwani, "Electroglottography and Vocal Fold Physiology," *Journal of Speech and Hearing Research*, vol. 33, pp. 245-254, 1990.

[21] K. E. Barner, "Nonlinear Estimation of DEGG Signals with Applications to Speech Pitch Detection," Applied Science and Engineering Laboratories, University of Delaware / A.I. duPont Institute, Wilmington, 1989.

[22] A. Camacho, 2008. [Online]. Available: http://www.cise.ufl.edu/~acamacho/english/curriculum.html. [Accessed 10 August 2016].

APPENDICES

Appendix A

Table 15: Example of a Subject's Calibration Matrix (Subject 31)

-3	-16.6824	-20.0602	-4.35249	11.6015	23.0475	1.22259	-1.5
-2.9	-14.7079	-23.0593	-2.3306	9.20218	24.0186	2.1338	-1.425
-2.8	-13.9314	-24.063	-1.66034	8.89051	24.2374	2.42679	-1.35
-2.7	-13.5647	-24.6634	-1.07981	8.68888	24.4729	2.64504	-1.275
-2.6	-13.2617	-25.0426	-0.543167	8.35761	24.6937	2.80602	-1.2
-2.5	-12.9541	-25.2199	-0.0760245	8.10937	24.8427	2.94148	-1.125
-2.4	-12.755	-25.4253	0.361305	7.83351	24.9835	3.05558	-1.05
-2.3	-12.5918	-25.6339	0.739931	7.59013	25.1735	3.17245	-0.975
-2.2	-12.4546	-25.8535	1.02037	7.3897	25.3355	3.31337	-0.9
-2.1	-12.3077	-26.0424	1.306	7.21519	25.4977	3.40425	-0.825
-2	-12.1847	-26.1946	1.65157	7.10404	25.6141	3.51842	-0.75
-1.9	-12.0901	-26.3948	1.84519	6.94963	25.7434	3.6348	-0.675
-1.8	-11.9698	-26.5831	2.03025	6.78497	25.8654	3.74478	-0.6
-1.7	-11.8889	-26.7149	2.24408	6.62147	25.999	3.84922	-0.525
-1.6	-11.7911	-26.8581	2.47787	6.46879	26.1162	3.99472	-0.45
-1.5	-11.6932	-26.9998	2.64043	6.33297	26.222	4.07809	-0.375
-1.4	-11.6058	-27.1654	2.83562	6.19192	26.3143	4.17905	-0.3
-1.3	-11.5102	-27.3277	2.95683	6.03648	26.4122	4.26899	-0.225
-1.2	-11.4192	-27.4857	3.06589	5.92279	26.5301	4.33662	-0.15
-1.1	-11.318	-27.6265	3.19612	5.79801	26.6494	4.44756	-0.075
-1	-11.2019	-27.7431	3.33476	5.6753	26.7685	4.53177	0
-0.9	-11.1215	-27.8832	3.42383	5.55196	26.8783	4.6394	0.075
-0.8	-11.0392	-27.9861	3.56751	5.46356	27.0126	4.73084	0.15
-0.7	-10.9301	-28.0971	3.69041	5.35848	27.1172	4.82328	0.225
-0.6	-10.824	-28.2374	3.82693	5.2394	27.24	4.90527	0.3
-0.5	-10.727	-28.3689	3.94514	5.13498	27.3603	5.03246	0.375
-0.4	-10.6192	-28.5157	4.06541	5.00998	27.4687	5.13612	0.45
-0.3	-10.5043	-28.7007	4.19735	4.91027	27.5788	5.21559	0.525
-0.2	-10.4256	-28.8868	4.30801	4.79931	27.7229	5.29107	0.6
-0.1	-10.329	-29.0942	4.44181	4.66271	27.8574	5.37318	0.675
0	-10.2297	-29.2883	4.52741	4.52741	27.9412	5.46787	0.75
0.1	-10.1306	-29.4906	4.66271	4.44181	28.0318	5.56005	0.825
0.2	-10.0477	-29.6516	4.79931	4.30801	28.1391	5.67702	0.9
0.3	-9.95317	-29.8896	4.91027	4.19735	28.2793	5.77975	0.975
0.4	-9.84356	-30.0739	5.00998	4.06541	28.4088	5.83661	1.05
0.5	-9.76322	-30.2446	5.13498	3.94514	28.5496	5.91313	1.125

0.6	-9.67062	-30.4142	5.2394	3.82693	28.6518	5.99312	1.2
0.7	-9.55487	-30.5855	5.35848	3.69041	28.7787	6.06168	1.275
0.8	-9.45886	-30.8519	5.46356	3.56751	28.9309	6.1551	1.35
0.9	-9.35517	-31.0672	5.55196	3.42383	29.1213	6.23035	1.425
1	-9.23489	-31.2613	5.6753	3.33476	29.2795	6.30942	1.5
1.1	-9.09409	-31.5272	5.79801	3.19612	29.4406	6.37061	1.575
1.2	-8.96802	-31.8222	5.92279	3.06589	29.627	6.45513	1.65
1.3	-8.84331	-32.0367	6.03648	2.95683	29.8142	6.53693	1.725
1.4	-8.71411	-32.3005	6.19192	2.83562	29.9816	6.59372	1.8
1.5	-8.57165	-32.617	6.33297	2.64043	30.1416	6.67772	1.875
1.6	-8.41441	-32.9165	6.46879	2.47787	30.309	6.76435	1.95
1.7	-8.26827	-33.3189	6.62147	2.24408	30.5051	6.83678	2.025
1.8	-8.12329	-33.6647	6.78497	2.03025	30.6656	6.93181	2.1
1.9	-7.98632	-33.9232	6.94963	1.84519	30.9144	7.0086	2.175
2	-7.85763	-34.2479	7.10404	1.65157	31.08	7.09854	2.25
2.1	-7.71593	-34.6484	7.21519	1.306	31.2839	7.20905	2.325
2.2	-7.58177	-35.1926	7.3897	1.02037	31.5286	7.32806	2.4
2.3	-7.40698	-35.6677	7.59013	0.739931	31.8021	7.45732	2.475
2.4	-7.20144	-36.2253	7.83351	0.361305	32.0463	7.5844	2.55
2.5	-7.01321	-36.983	8.10937	-0.0760245	32.237	7.72747	2.625
2.6	-6.83041	-37.8266	8.35761	-0.543167	32.4961	7.8385	2.7
2.7	-6.6465	-38.5247	8.68888	-1.07981	32.749	8.0124	2.775
2.8	-6.4503	-39.9049	8.89051	-1.66034	33.0646	8.20988	2.85
2.9	-6.09905	-41.2503	9.20218	-2.3306	33.423	8.46431	2.925
3	-5.48889	-44.7247	11.6015	-4.35249	36.1093	9.4782	3

Appendix B

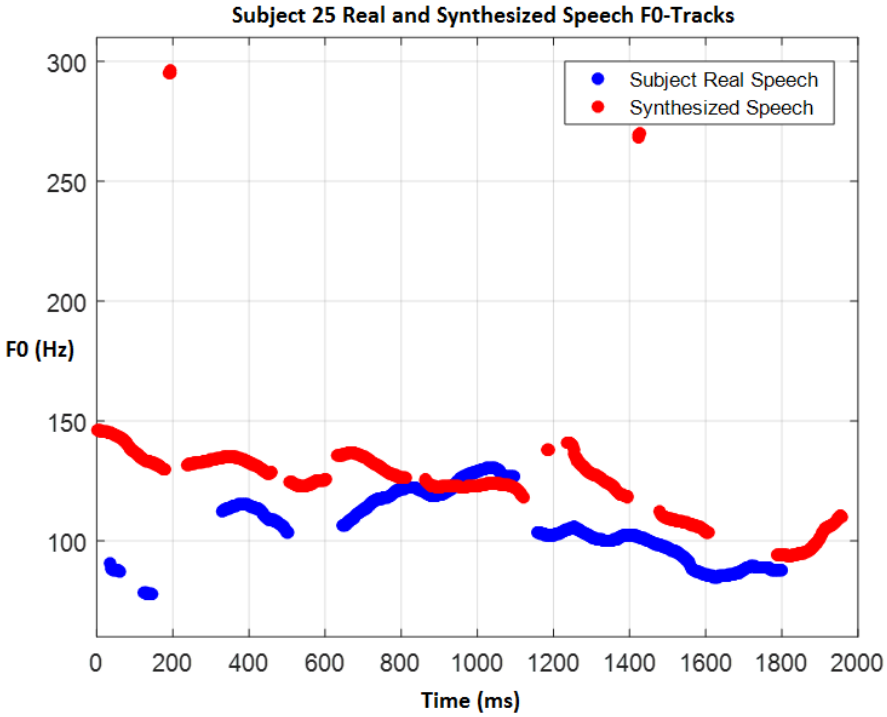

Figure 39: F0-Track of Subject 25's Real and Synthesized Speech

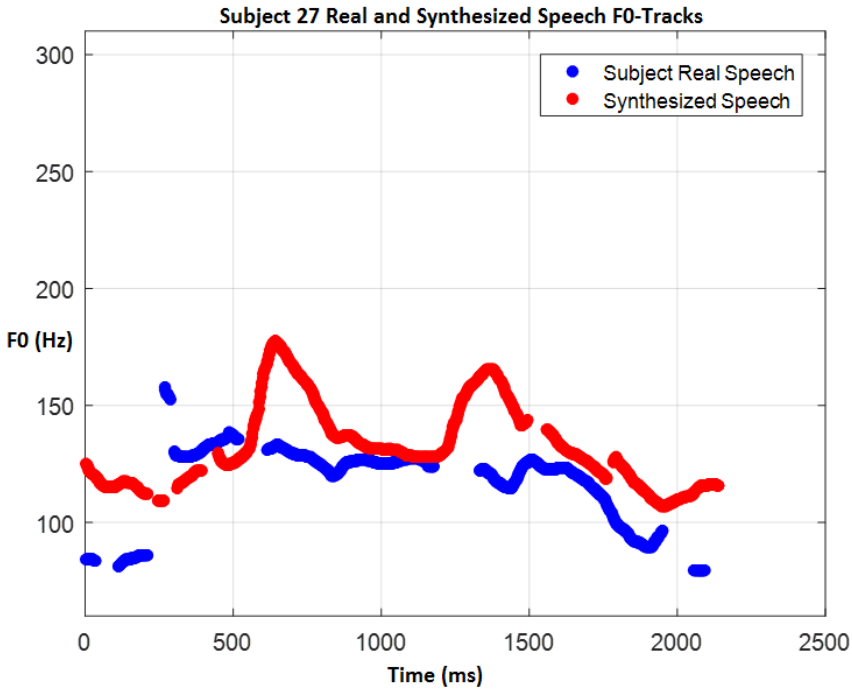

Figure 40: F0-Track of Subject 27's Real and Synthesized Speech

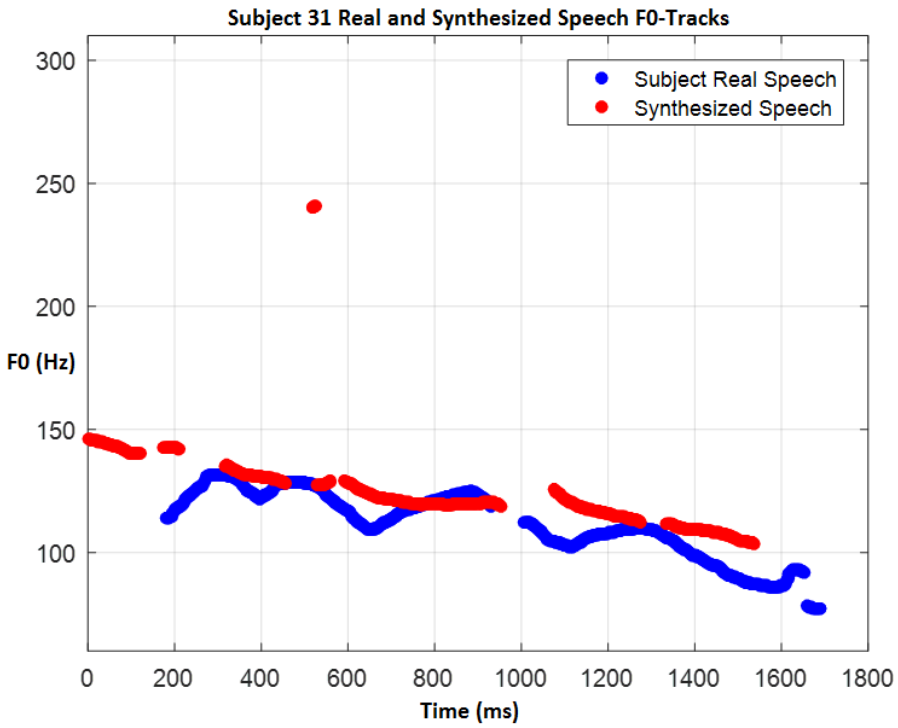

Figure 41: F0-Track of Subject 31's Real and Synthesized Speech

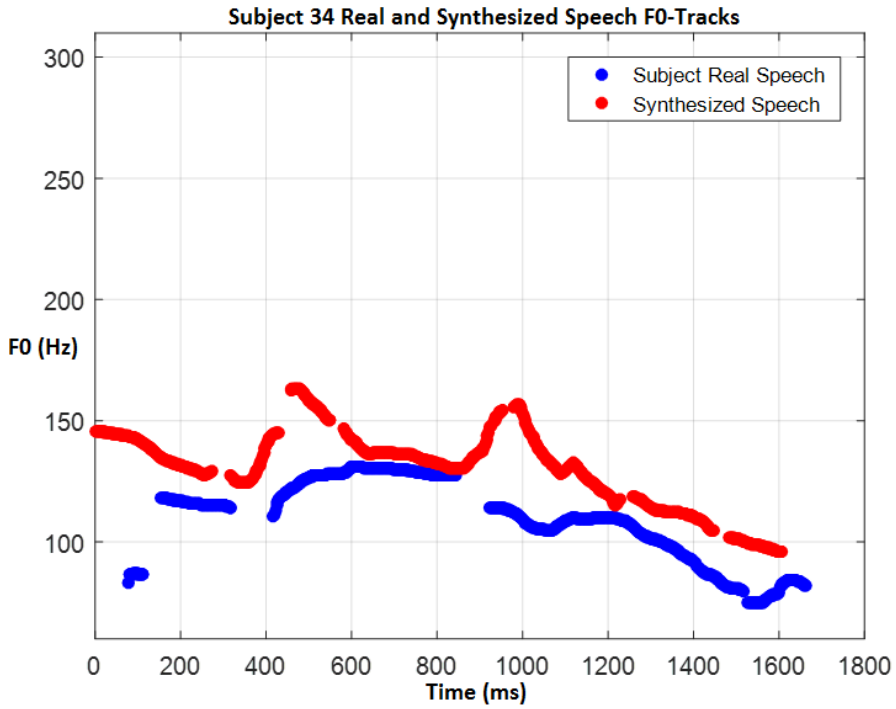

Figure 42: F0-Track of Subject 34's Real and Synthesized Speech

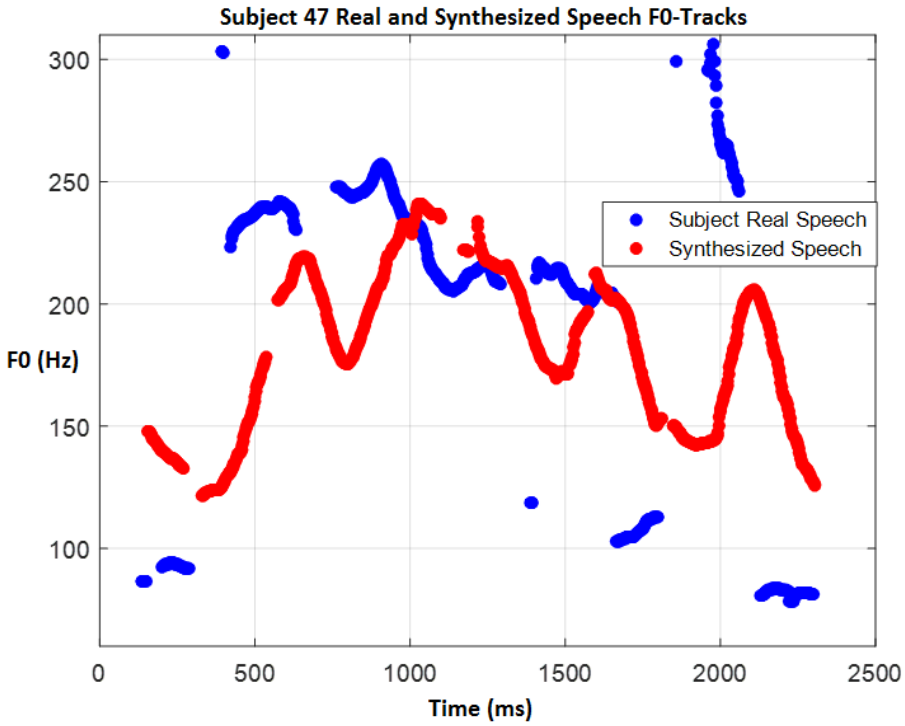

Figure 43: F0-Track of Subject 47's Real and Synthesized Speech

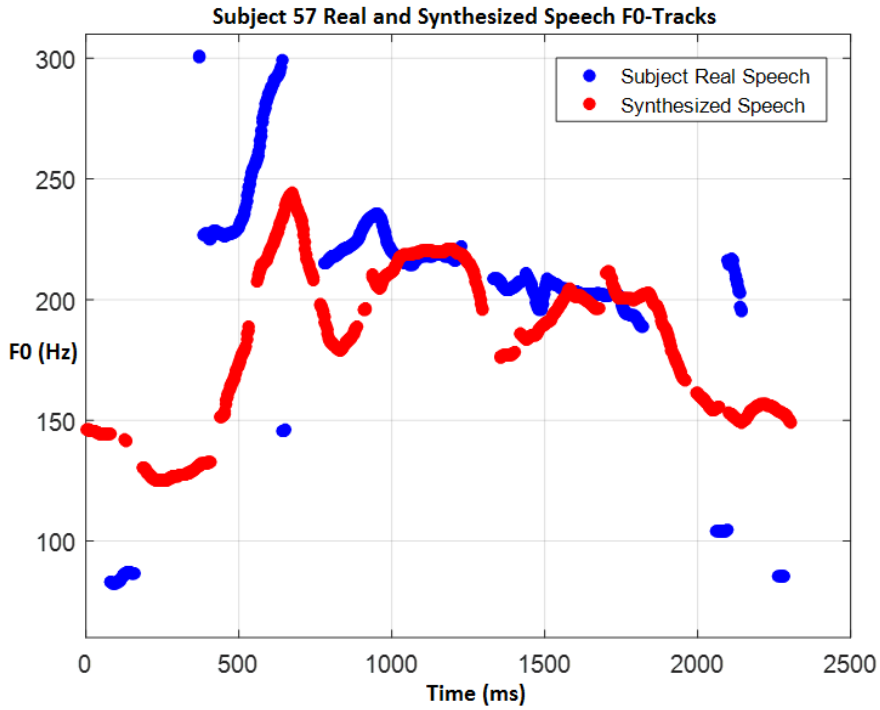

Figure 44: F0-Track of Subject 57's Real and Synthesized Speech