Least-Squares Mapping from Kinematic Data to Acoustic Synthesis Parameters for Rehabilitative Acoustic Learning

Xiangyu Zhou
Marquette University

LEAST-SQUARES MAPPING FROM KINEMATIC DATA TO ACOUSTIC SYNTHESIS PARAMETERS FOR REHABILITATIVE ACOUSTIC LEARNING

By

Xiangyu Zhou, 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 2016

ABSTRACT

LEAST-SQUARES MAPPING FROM KINEMATC DATA TO ACOUSTIC SYNTHESIS PARAMETERS FOR REHABILITATIVE ACOUSTIC LEARNING

Xiangyu Zhou

Marquette University, 2015

Thousands of people suffer from dysarthria resulting from neurological injury of the motor component of the motor-speech system, and need to rely on alternative methods to communicate in daily life, such as body language or text-to-speech [1] . However, there are currently very few effective rehabilitative therapies for helping these patients improve their speech. Because of this, research is needed to develop better rehabilitative therapies. One such area of research is the use of involuntary acoustic learning. The Speech and Swallowing lab at Marquette University has an Electromagnetic Articulography (EMA) system to collect kinematic data and a software system called Rehabilitative Articulatory Speech Synthesizer (RASS) that is able to create the necessary synthesized acoustic feedback to study the effects of these kind of therapies.

One key aspect of the RASS system is the mapping from kinematic sensor data to acoustic synthesis parameters. This is a complex problem that depends on individual subject anatomy and vocal tract patterns. Currently, the RASS system uses a simple piecewise linear method, but it would be advantageous to improve this to be more accurate across a wider range of vocal configurations. The goal of the research work presented here is to develop and test new approaches for kinematic to synthesis mapping, in the hopes of improving the quality and intelligibility of the RASS system.

Results indicate that the new mapping gives reduced mapping error. Ultimately, the impact of this work is that it provides researchers with a more accurate method for mapping kinematic data to synthesis parameters.

ACKNOWLEDGEMENT

Xiangyu Zhou

I would like to express my gratitude to the people who were my guides, my friends, and my support during this journey through Master program.

I would also like to express my sincere gratitude and appreciation to my advisor, Prof. Michael T Johnson, for his supervision, and for the valuable knowledge that he shared with me. I have learned so much from his wisdom, carefulness, and visions.

In addition, I would like to express my sincerely appreciation to my committee members, Prof. Jeffrey Berry and Prof. Richard Povinelli for their time support my work and dedicated to examine and review my work and also for their valuable feedback, as well as useful suggestions that helped to make this thesis a more complete document.

There are people in Marquette Speech and Signal Processing lab whose kindness and help are immeasurable. I wish to express special thanks to my friends, An Ji, Jianglin, Wang and Andrew Kolb for their enjoyable discussions, valuable comments and assistance in my research work.

Finally, I would like thank my family. Even though I'm geographically separated from them, they have always encouraged me, supported me and believed in me through these past few years. Without their support and love, I would not have finished this thesis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 Introduction

1.1. Introduction

Thousands of people suffer from dysarthria resulting from neurological injury of the motor component of the motor-speech system, and need to rely on alternative methods to communicate in daily life, such as body language or text-to-speech [1]. However, there are currently very few effective rehabilitative therapies for helping these patients improve their speech. Because of this, research is needed to develop better rehabilitative therapies. One such area of research is the use of involuntary acoustic learning. For example, record the articulatory kinematics of dysarthric speakers, and using modified kinematic-driven acoustic feedback， we can create involuntary sensory-motor learning. The Speech and Swallowing lab at Marquette University has an Electromagnetic Articulography (EMA) system to collect kinematic data and a software system called Rehabilitative Articulatory Speech Synthesizer (RASS) that is able to create the necessary synthesized acoustic feedback to study the effects of these kind of therapies.

One key aspect of the RASS system is the mapping from kinematic sensor data to acoustic synthesis parameters. This is a complex problem that depends on individual subject anatomy and vocal tract patterns. Currently, the RASS system uses a simple piecewise linear method, but it would be advantageous to improve this to be more accurate across a wider range of vocal configurations. The goal of the research work

presented here is to develop and test new approaches for kinematic to synthesis mapping, in the hopes of improving the quality and intelligibility of the RASS system.

1.2. Electromagnetic Articulography

The RASS system is based on using Electromagnetic Articulography (EMA) to collect the kinematic movement data of sensors placed on the tongue, lips and jaw of human subjects. The EMA system software includes automated correction of head movement so that the data is with respect to the subject, with average sensor tracking errors below 0.5mm for dynamic tracking [2].The data is then additionally processed to reference it to each individual subjects' articulatory space, using a calibration process based on biteplate data as described in more detail in Section 2.2.4. Acoustic data is also collected through the EMA system and synchronized to the kinematic data, although it is not necessary for use of the RASS system.

1.3. RASS

The Rehabilitative Articulatory Speech Synthesizer (RASS) used in the Speech and Swallowing lab at Marquette University is a real time system which perturbs the acoustic data in the formant space and returns the perturbed speech back as biofeedback to the subject in real time. It receives streaming data from the EMA system and interacts with the experimenter and the real time VTDemo system (Vocal Tract Acoustics Demonstrator) which synthesizes the acoustics. The RASS system is described in more detail in Section 2.1.

The VTDemo system is an interactive articulatory synthesizer, originally created by Mark Huckvale [2], based on the program VTCALCS [2] by Satrajit Ghosh at Boston University. The synthesizer model itself is based on the work of Shinji Maeda [3]. VTDemo takes a set of seven vocal tract shape parameters taken from Maeda's work and converts them into a vocal tract area function which is then used to filter a voicing signal form a modeled voice source and allow real-time synthesis of the acoustic signal as the articulatory parameters.

In order to synthesize an acoustic signal, the VTDemo synthesis component first requires these seven synthesis parameters as input. These parameters must be estimated from the kinematic data of the subject in order to connect the EMA system and the synthesis system. The current system uses kinematic data from four sample vowels as well as the overall dynamic range of motion, and creates a simple three-segment piecewise linear mapping that maps selected sensor positions to individual synthesis parameter values. This is an approximate process and the mapping must often be hand adjusted to be useable in practice.

Once the speech is synthesized, the Audapt software system adjusts the formant values as desired for a particular subjective experiment, and resynthesizes the data a second time. This allows direct control over the acoustics of patient's speech in real time. In rehabilitative studies, this modified signal is returned to the subjects through headphones and the resulting involuntary change to the subject's acoustic response is recorded for analysis. This experimental process is shown in Figure 1 below.

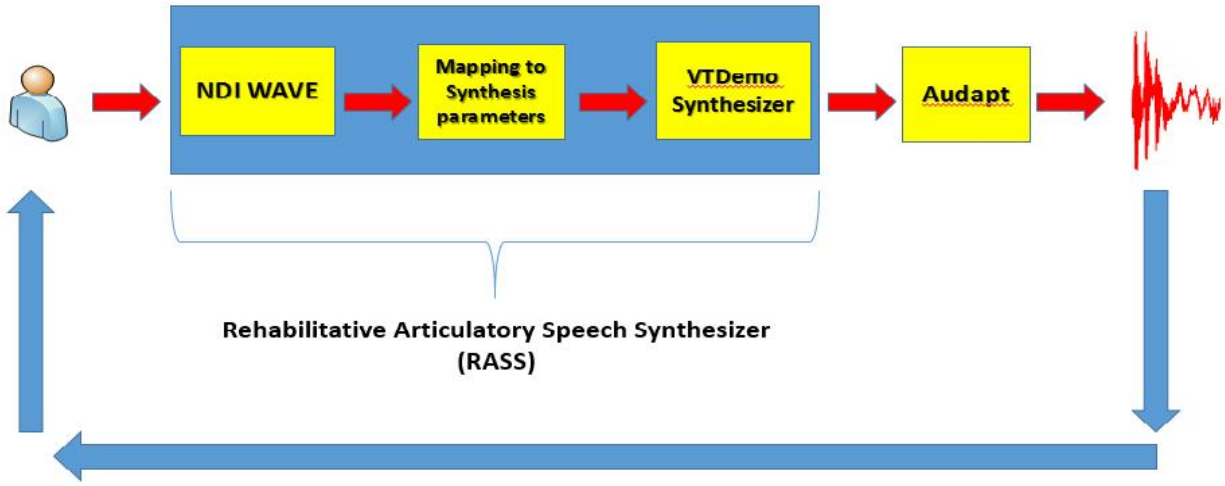

Figure 1: The RASS system

1.4. Proposed Approach

The goal of this work is to create an improved kinematic-to-synthesis parameter mapping method to connect the kinematic EMA system to the RASS synthesizer. The proposed approach for this is based on a least-squares mapping concept where training data from the subject is used to build a multiple-input multiple-output mapping model.

The first step of the new approach is to develop higher quality articulatory features from the raw kinematic data. To do this, a mesh model of the subject's palate is estimated from palate trace data collected for each subject. This allows for the computation of sensor-to-palate distances, giving articulatory features that directly relate to vocal tract opening size.

Using these improved articulatory features, calibration data from multiple vowels and consonants is used to create a data matrix representing the relationship between

articulatory features and acoustics. This is then input to a least-squares algorithm to build a linear regression model, which maps input values of articulatory features to output synthesis parameters across the entire acoustic space.

The new approach is compared to several prior mapping approaches using mean square error method. For evaluating accuracy and effectiveness, the experimental work is divided into two parts, single phoneme reconstruction and sentence reconstruction. To assess the system, articulatory features are computed for individual phonemes, and then mapped to synthesis parameters through the linear mapping equation. Mean square error between synthesis parameters from the new approach and from previous approaches can then be directly compared. In addition to the mean square error, we also compare the acoustic results between new and prior approaches using Perceptual Evaluation of Speech Quality (PESQ) standard.

1.5. Research Objectives

Through the proposed kinematic mapping structure, we can re-express participant's articulatory movements and use them to support the study of sensorimotor relationships and further our understanding of feedforward and feedback mechanisms in speech motor control [4] . How to process these kinematic data and relate them to the needed synthesis parameters is one of the key steps for the RASS system. In this thesis, we introduce a new method for the mapping between kinematic data and synthesis parameters by using articulatory features with the three dimensional virtual

vocal tract. The new method is compared to several prior methods to show the benefits of the new approach. Overall, the new mapping method has the ability to create an accurate linear relationship between kinematic speech data and synthesis parameters.

1.6. Overview of Thesis

The remainder of this thesis is organized into the following sections: Introduction (Chapter one), Background (Chapter two), Palate Mesh creation (Chapter three), Kinematic to Synthesis Parameter Mapping (Chapter four), and Conclusion (Chapter Five).

CHAPTER 2 Background

In this chapter, fundamental concepts and technical background that relate to the articulatory to synthesis mapping problem will be introduced. This includes an overview of the RASS system and its components, details of the EMA system itself, the underlying model of the VTDemo synthesizer, and a brief overview of the Audapt acoustic modification framework. In addition, prior methods used in RASS for articulatory to synthesis mapping will be reviewed in detail.

2.1. Overview

Articulatory models describe the vocal tract shape by means of a small number of control parameters [5]. Synthesizer parameters may be determined values that represent vocal tract function to generate corresponding acoustic waveforms. In the Rehabilitative Articulatory Speech Synthesizer (RASS) system, articulatory variables computed from sensor position describe the shape of the vocal tract which are translated to synthesis parameters to control acoustic feedback. Figure 1 in Chapter 1 illustrates the three components of this, which include the NDI Wave system, the mapping system which maps kinematic to synthesis parameters, and the VTDemo system. The NDI Wave system is used to collect kinematic data in real time, while the mapping system maps the kinematic data to synthesis parameters. After mapping, the generated synthesis parameters are used as inputs to the VTdemo system to generate an acoustic waveform, which is then modified by Audapt software and returned to the subject via headphone.

Figure 2 shows the detailed functional decomposition of the RASS system and its functionality.

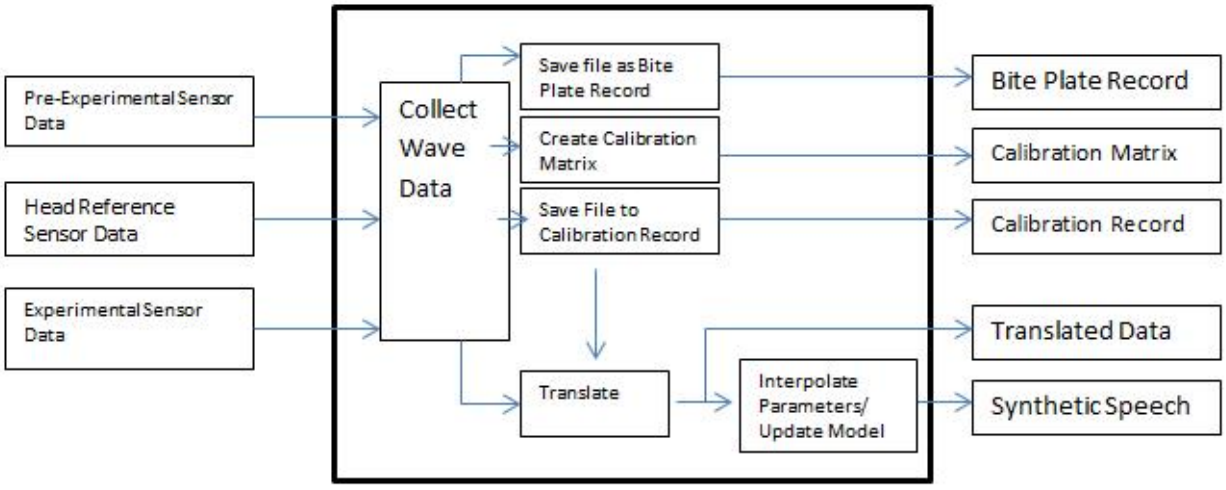

Figure 2: Software structure of the RASS system

2.2. NDI Wave system and Data collection

The Northern Digital (NDI) Wave system is used to collect kinematic speech position data. Data records typically include a bite-plate record, a palate trace record and multiple speech records. For each of these records, the subject wears “orientation” glasses that have a single 6DOF (degree of freedom) reference sensor for the purpose of head correction and coordinate space translation.

2.2.1 NDI Wave system

The NDI Wave system is an electromagnetic articulography (EMA) speech research system which tracks real-time articulatory orofacial movements and kinematics [6]. It includes:

- Field Generator
- Mounting arm
- System Control Unit
- System Interface Units
- Microsensors
- Reference Sensor
- WaveFront™ data collection and real-time viewing software, with audio signal synchronization functionality
- Audio synchronization cable
- Palate probe

The NDI Wave system supports three dimensional (3D) tracking of 5 or 6 degree-of-freedom (5-DOF, 6DOF) sensors in a static electromagnetic field. This is based on the basic principle of two-dimensional magnetometer systems [7]. Through a static field generator, a signal can be induced in sensors via electromagnetic induction. Through this the sensor position and orientation can be captured. Figure 3 shows the layout of NDI Wave system.

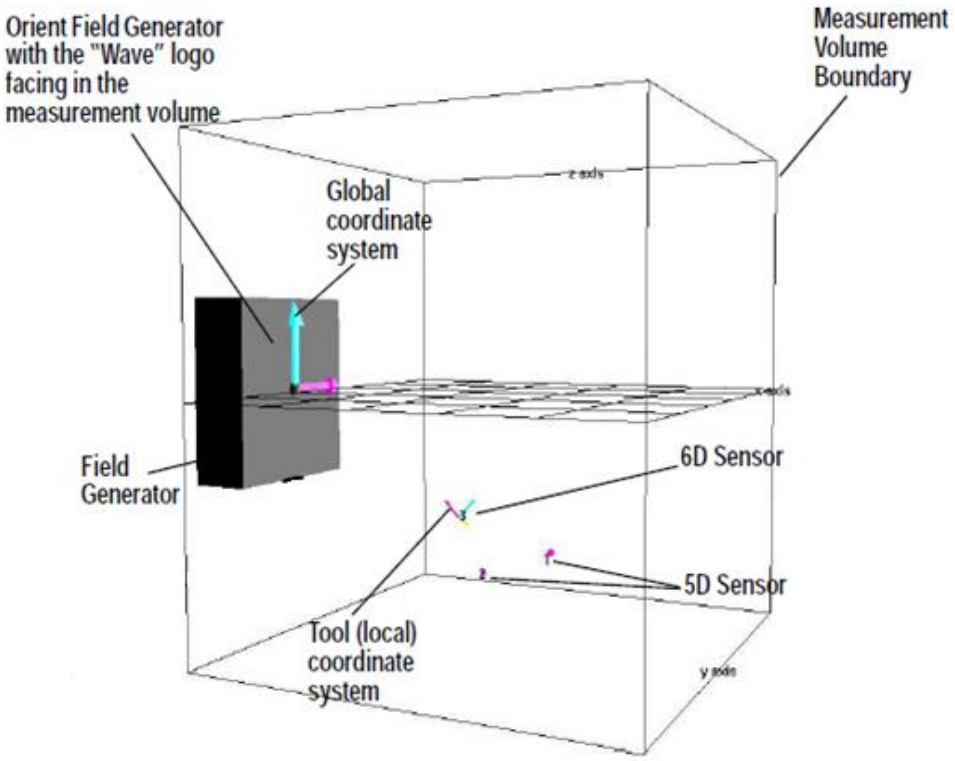

Figure 3: Layout of the NDI Wave system

2.2.2 Sensor placement

A typical configuration for the RASS system includes 6 articulatory sensors, one 6 degree-of-freedom sensor and five 5 degree-of-freedom sensors. Figure 4 shows an example of these sensors placed on a human subject and Figure 5 shows the sensor placement inside subject's mouth.

Figure 4: Sensor placement on a human subject

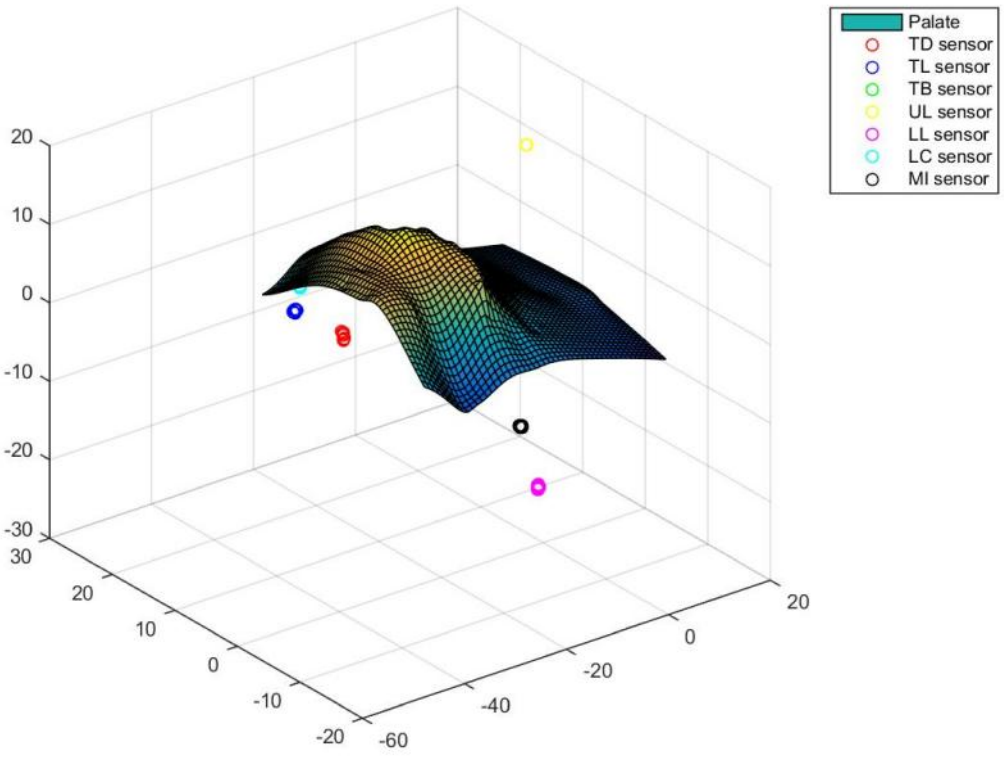

Figure 5: 3D sensor placement

Specific placement includes:

1. **REF** (not shown): Reference sensor, 6-DOF, placed on a plastic glasses frame approximately superficial to the intersection of the superior aspect of the nasal bone and the glabella of the frontal skull bone. The purpose of REF is for head correction and coordinate space translation;
2. **MI**: Mandibular Incisor sensor, 5-DOF sensor, placed at the intersection of the central mandibular incisors (labial surface), abutting the enamel-gingival border;
3. **UL**: Upper lip sensor, 5-DOF sensor, placed midsagittally at the intersection of the inferior aspect of the philtrum and the vermillion border;
4. **LL**: Lower lip sensor, 5-DOF sensor, placed midsagittally along the vermillion border of the lower lip;
5. **TB**: Tongue blade sensor, 5-DOF sensor, placed midsagittally along the dorsal surface of the apex of the tongue, approximately 5 mm posterior to the tongue tip;
6. **TD**: Tongue dorsum sensor, 5-DOF sensor, placed midsagittally along the posterior tongue dorsum approximately 40 mm posterior to the tongue tip.

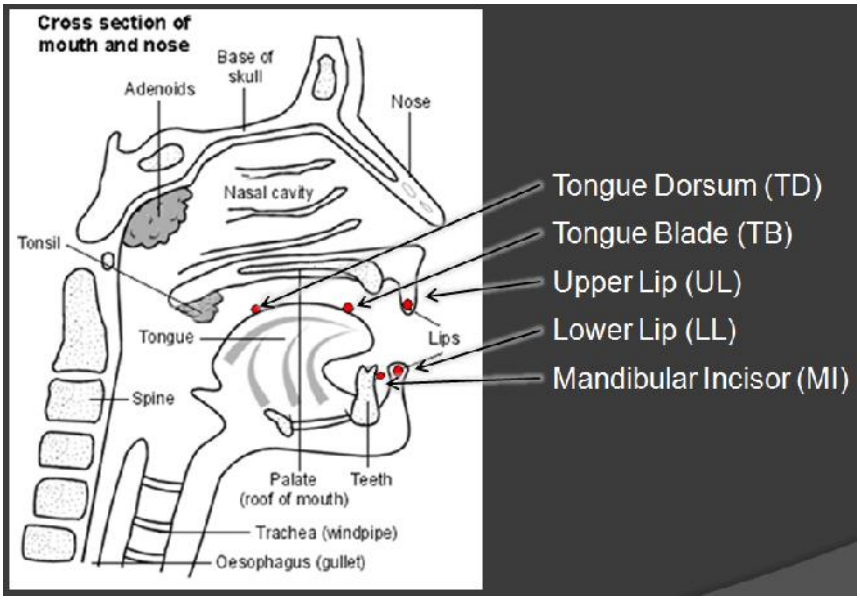

Figure 6: Sensor position side view

2.2.3 Bite-plate record

The bite-plate record is used to gather the data needed to determine each subjects' personal coordinate system, where the x axis lies along the juncture of the midsagittal and maxillary occlusal plane, the y axis runs vertically perpendicular (upwards), and the z axis runs horizontally perpendicular (to the subject's left). Two sensors are placed on the bite-plate: one at the maxillary central incisors (OS) and one along the midsagittal plane at the bisection between the back molars (MS) as shown in Figure 7.

Figure 7: MS and OS sensor positions in the bite-plate

Figure 8 shows the placement of the bite-plate within the subject's mouth.

Figure 8: Illustration of bite plate record

A bite-plate calibration algorithm, described in Section 2.2.3, is used to translate and rotate the subjects' individual coordinate system so that the origin is at the OS sensor and the articulatory space is referenced to the midsagittal and maxillary occlusal planes.

2.2.4 Calibration process

After the data have been recorded, they are in the local coordinate space defined by the head reference sensor. In order to create a meaningful articulatory working space, the data have to be calibrated into a normalized articulatory space. The calibration process includes bite-plate calibration, offset adjustment and final verification. The baseline of articulatory space is based on each subject's anatomy, as shown in Figure 9.

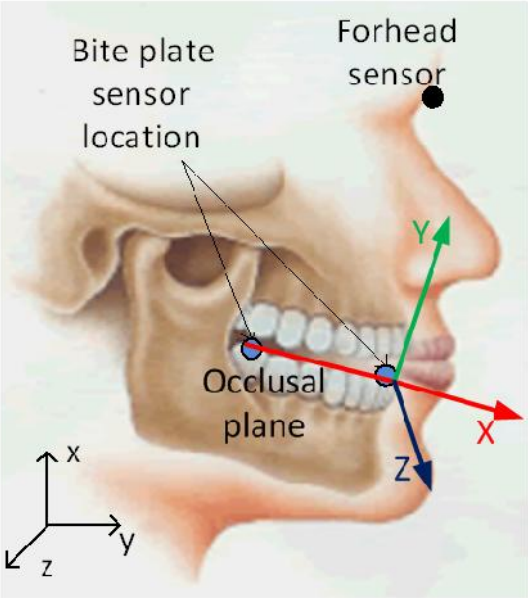

Figure 9: Target coordinate system

2.2.4.1. Bite-plate calibration process

Because the OS, MS, and REF sensors are placed within the mid-sagittal and maxillary occlusal planes, the relative positioning within the bite-plate record can be implemented to transform the data into the desired articulatory working space.

The articulatory working space is based on individual's anatomy structure, as shown Figure 9. The origin of the coordinate system is set as the central point of the upper maxillary incisors. The vertical plane is set as the mid-sagittal plane, and the horizontal plane is defined as the maxillary occlusal plane, which is the plane of contact between the maxillary and mandibular teeth. Relative to these two coordinate planes, the X axis indicates forward and backward movement, the Y axis expresses upward and down ward movement, and the Z axis represents lateral movement. Thus, the mid-sagittal plane is given by the X-Y axes and the maxillary occlusal plane by the X-Z axes.

The positive x axis is forward of the incisors, so that the negative x axis follows the mid sagittal line of the occlusal plane toward the back of the throat. The positive z axis runs perpendicularly to the x axis on the occlusal plane toward the subject's right. The positive y axis is perpendicular to the occlusal plane in the upward direction. This convention follows the "right hand rule," with the origin at the maxillary central incisors.

The fundamental goal of the data calibration process, called bite-plate calibration, is to ensure that the coordinate system represented by the data follows as closely as possible to the theoretical target articulatory working space mentioned above.

When the bite-plate calibration has been applied, the REF, OS, and MS sensors create a coordinate articulatory working space. OS is located at original point [0, 0, 0], MS is on the x-axis, and REF is in the mid sagittal plane.

However, there is one final adjustment to be made to this working space. Due to the width of the sensors themselves (and that of the incisors), the center of the OS sensor is

not exactly at the central tip of the upper maxillary incisors. In addition, because the incisors typically bite into the dental wax on the bite place down to the level of the tongue depressor surface, both the OS and MS sensors are placed in the dental wax at a depth such that the center line of these two sensors is slightly above the tip of the upper maxillary incisors. To compensate for this offset, a final translation can be done equal to the expected distance from the center of the OS sensor to the true tip of the upper maxillary incisors. On average, this is about a -4mm offset horizontally (negative meaning toward the posterior), and about a -1mm offset vertically (downward). Figure 10 shows the sensor measurement details.

Figure 10: Sensor measurement

2.2.4.2 Verification

After the bite plate calibration and offset adjustment, the system verifies the orientation and coordinate system. The verification includes the following items:

- The REF sensor has a relatively large positive y value. (Vertical orientation correct)

- The MS sensor has a relatively large negative x value (Horizontal orientation correct.)
- The LAT sensor (when present) has a relatively large positive z value (Lateral orientation correct.)
- The tongue blade sensor data is in front of the tongue dorsum sensor.
- The upper lip sensor is above the lower lip sensor.
- The tongue blade and dorsum sensors are generally behind both the lower lip and the upper lip sensors.
- The jaw sensor is generally below all other sensors.
- The palate data field is generally above all tongue sensors.
- The tongue lateral sensor (where one is used) has a generally large positive z value.
- The lip lateral sensor has a generally large positive z value.

2.2.5 Palate trace process

The palate trace process is used to record each subject's palate shape. Through this process, we can describe the oral position of the vocal tract through combining the position of palate trace and sensors inside the subject's mouth. This process includes both the surface of palate and the perimeter of the teeth. The sensor's trace is started at the central maxillary incisors, and the probe is swept straight back along the palate surface toward the uvula, then back and forth laterally from right to left dentition. Finally, the probe is swept around the lingual surface of the tips of the maxillary

dentition starting at subject's left-posterior and ending at right-posterior. Figure 11 shows the track of sensor applying palate trace in details.

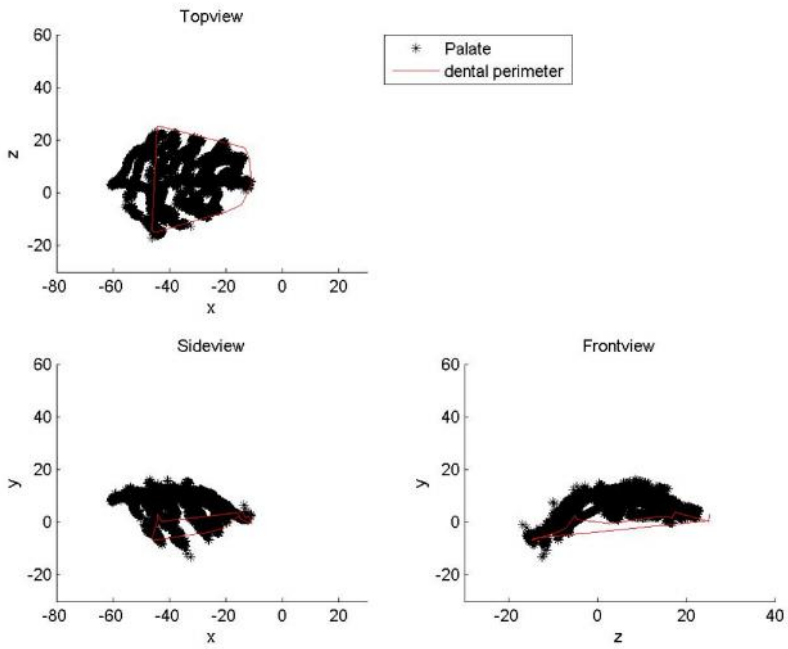

Figure 11: Typical palate trace data

2.2.5. Speech recording process

The Marquette Electromagnetic Articulography corpus of Mandarin-Accented English (EMAMAE) is used in this work [8]. The EMAMAE dataset includes 40 speakers, 20 native English speakers and 20 native Mandarin speakers, balanced equally between male and female speakers. The dataset was collected using NDI Wave system. Calibration data included a bite plate record and plate trace described in the previous section 2.2.3 and 2.2.4; Articulatory movement was recorded by sensors (MI, TB, TD, LC, TL, UL and LL). Speech information included sets of contrastive words as well as continuous speech.

2.3. VTDemo Synthesizer

2.3.1. Source-filter and Maeda model

Figure 12 describes the speech production process. Air from the lungs is forced through the trachea to the larynx and vocal folds. In voiced sounds such as vowels, these vocal folds are held closed with a certain tension, and the pressure of the air creates a quasi-periodic vibration of the vocal folds. This creates a glottal excitation signal which moves through the vocal tract articulators consisting of the throat, mouth, tongue, and lips, which together act as a filter to control the production of sound.

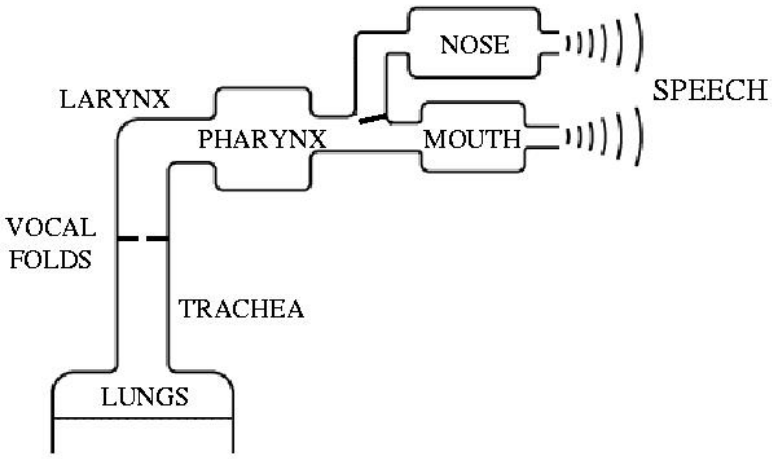

Figure 12: Source-filter model of the vocal tract

The VTDemo system is based on the work of Shinji Maeda [3] which models the vocal tract using a set of parameters based on principal-components analysis as shown in Figure 13

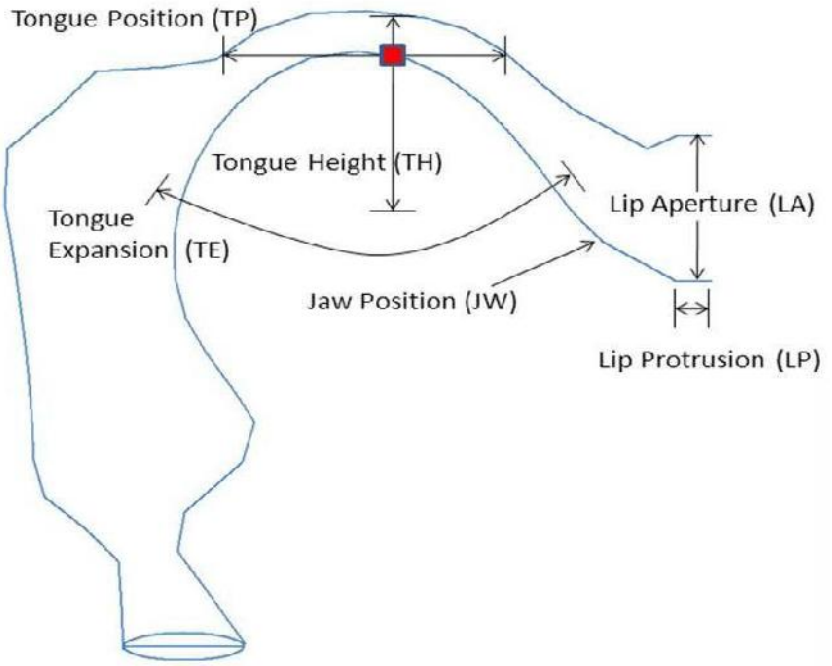

Figure 13: Maeda model

VTDemo takes a set of seven vocal tract shape parameters taken from Maeda's work and converts them into a vocal tract area function, which is then used to filter a voicing signal form a modeled voice source creating real-time synthesis of an acoustic signal. Figure 14 shows the interface of a stand-alone version of the VTDemo synthesizer.

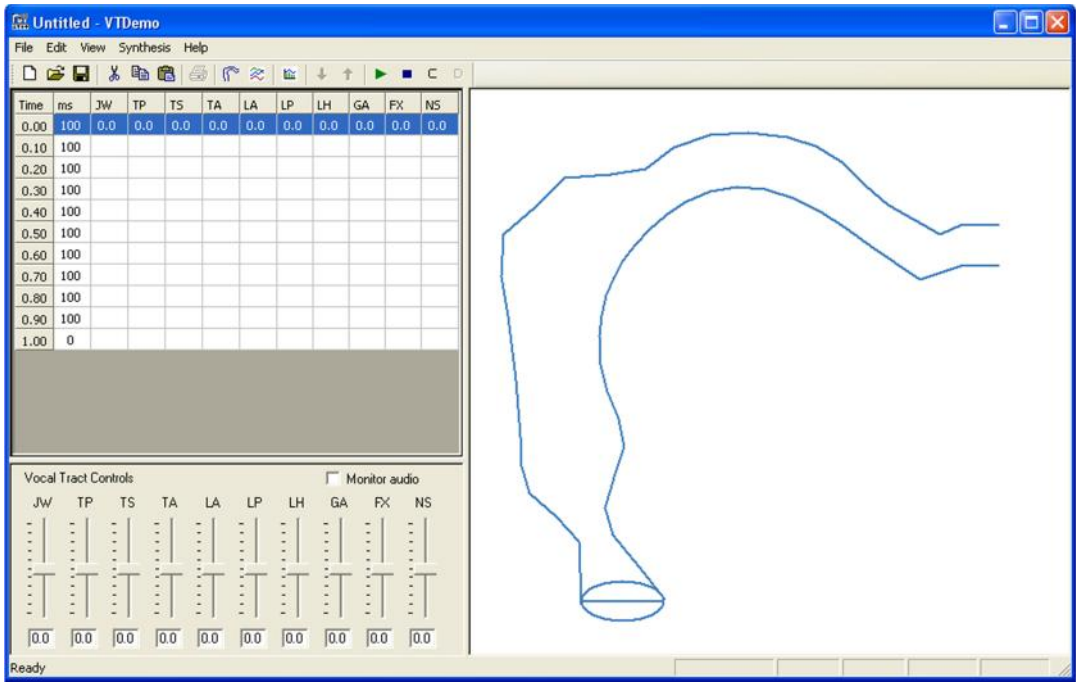

Figure 14: VTDemo software interface

The synthesis model is driven by the following acoustic synthesis parameters:

1. “JW” describes the jaw movement during the speech;
2. “TP” describes the expansion of tongue;
3. “TS” describes the height of tongue;
4. “TA” describes the status of tongue expansion;
5. “LA” describes the opening and closing status of lip;
6. “LP” describes the expansion of lip;
7. “LH” describes the vertical expansion of larynx
8. “GA” describes the opening and closing of glottis
9. “FX” describes fundamental frequency
10. “NS” describes the status of velopharyngeal port opening

2.4. Kinematic to Synthesis mapping

Previously, several different methods have been implemented in the RASS system for mapping kinematic data to synthesis parameters, including a 4-point piecewise linear mapping method, a 2-point piecewise linear mapping method, and a quantile based method. Each of these three share a common framework in terms of which synthesis parameters are controlled by which kinematic variables. Table 1 shows the details of synthesis parameter and their corresponding kinematic source for these approaches.

Table 1： Synthesis Parameters and Kinematic Data Sources as implemented by prior mapping methods

Synthesis Parameter	Variability	Kinematic Source
Segment Duration (ms)	Static	Sampling Rate of NDI-Wave
Jaw Position (JW)	Dynamic	Varies directly with MI_y position
Tongue Position (TP)	Dynamic	Varies inversely with mean of the TB_x and TD_x positions
Tongue Shape (TS)	Dynamic	Varies directly with the mean of the TB_y and TD_y positions
Tongue Expansion (TA)	Static	Set to neutral (0) value
Lip Aperture (LA)	Dynamic	Varies directly with the distance between UL_xy and LL_xy
Lip Protrusion (LP)	Dynamic	Varies directly with LL_x position
Larynx Height (LH)	Static	Unspecified
Glottal Aperture (GA)	Static	Unspecified
Fundamental Freq. (FX)	Static	Unspecified
Velopharyngeal Port Opening (NS)	Static	Unspecified

2.4.1. 4-point piecewise linear mapping method

The 4-point piecewise linear mapping directly maps kinematic data to synthesis parameters using 4 coordinate pairs based on subject dynamic range and subject vowel

space. The two end-points of the map are determined from the minima and maxima kinematic values of an extrema speech record taken from the subject, and the two internal points are determined from the average minima and maxima kinematic values across a set of four stationary vowel records. Details of this mapping are as follows:

Application sensors: TB, TD, MI, UL and LL

Synthesis parameters controlled: LP, LA, TH, TE, TP, and JW

Synth parameters fixed settings:

- Larynx Height (LH): Fixed at 0
- Glottal Aperture (GA): Fixed at continuous voicing
- Fundamental Frequency (FX): Fixed at 0 (default value)
- VP Opening (NS): Fixed at VP Closed

The specific articulatory mappings:

- The average Y position of the MI sensor is used to directly calculate the value MI_y which maps onto the JW parameter.
- The average X positions of the TB and TD sensors is the value TBD_x which inversely maps onto the TP parameter
- The average Y positions of the TB and TD sensors is the value TBD_y which maps onto the TS parameter
- The average Y positions of the TB and TD sensors is the value TBD_y which inversely maps onto the TA parameter

- The Euclidean distance between the UL and LL sensors is the value UL_LL which maps onto the LA parameter
- The X position of the LL sensor is used to directly calculate the value LL_x which maps onto the LP parameter

The linear mapping for the JW parameter is shown in Figure 15 with specific coordinates as follows:

- Extreme min value of MI_y correspond to -3 (MI_y _ext_min,-3)
- Mean minimum value of vowel records correspond to -1.5 (MI_y _vowel_min, -1.5)
- Mean maximum value of vowel records correspond to 0.5 (MI_y _vowel_max, 0.5)
- Extreme max value of MI_y correspond to +3 (MI_y _ext_max,+3)

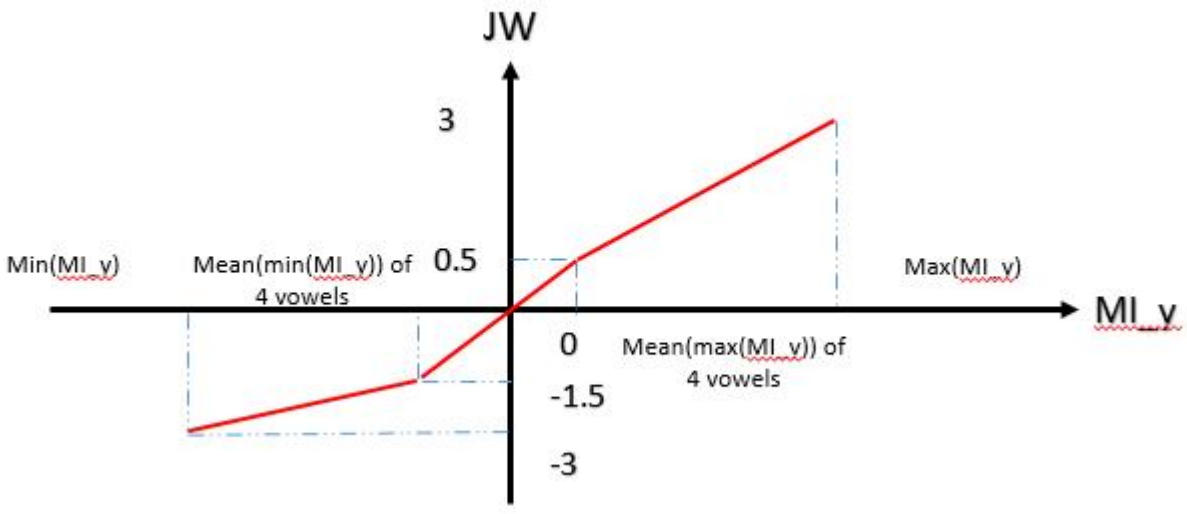

Figure 15: The 4 point piecewise mapping for the JW parameter

The linear mapping for the TP parameter is shown in Figure 16 with specific coordinates as follows:

- Extreme min value of TBD_x correspond to 3; (TBD_x _ext_min,3)
- Mean minimum value of vowel records correspond to 2 (TBD_x _vowel_min, 2)
- Mean maximum value of vowel records correspond to -2(TBD_x _vowel_max, -2)
- Extreme max value of TBD_x correspond to 3 (TBD_x _ext_max,-3)

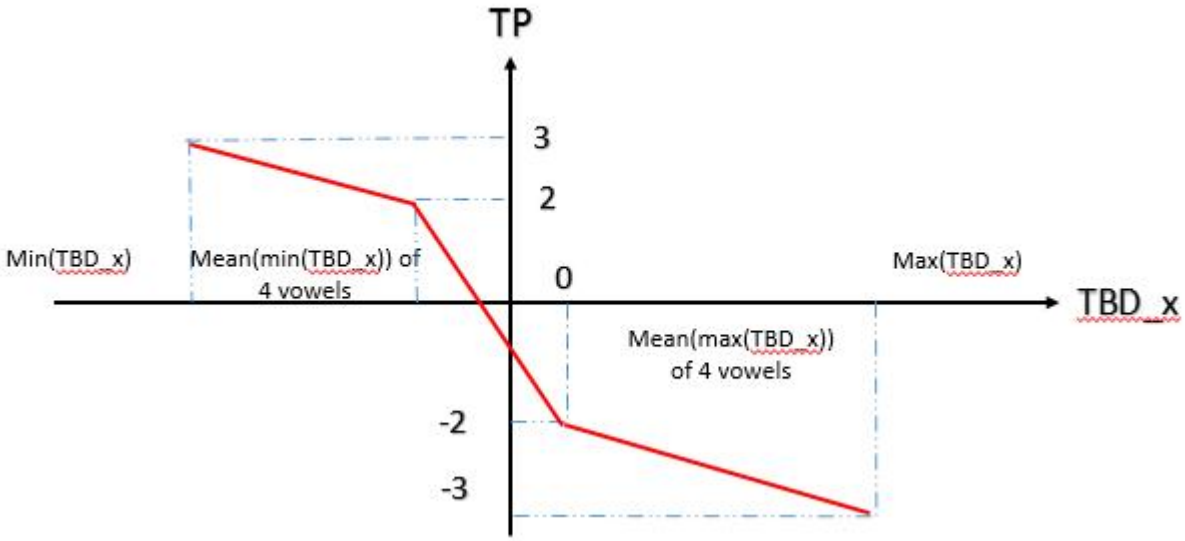

Figure 16: The 4 point piecewise mapping for the TP parameter

The linear mapping for the TS parameter is shown in Figure 17 with specific coordinates as follows:

- Extreme min value of TBD_y correspond to -3; (TBD_y _ext_min,-3)
- Mean minimum value of vowel records correspond to 0 (TBD_y _vowel_min, 0)
- Mean maximum value of vowel records correspond to 1 (TBD_y _vowel_max, 1)
- Extreme max value of TBD_y correspond to +3 (TBD_y _ext_max,+3)

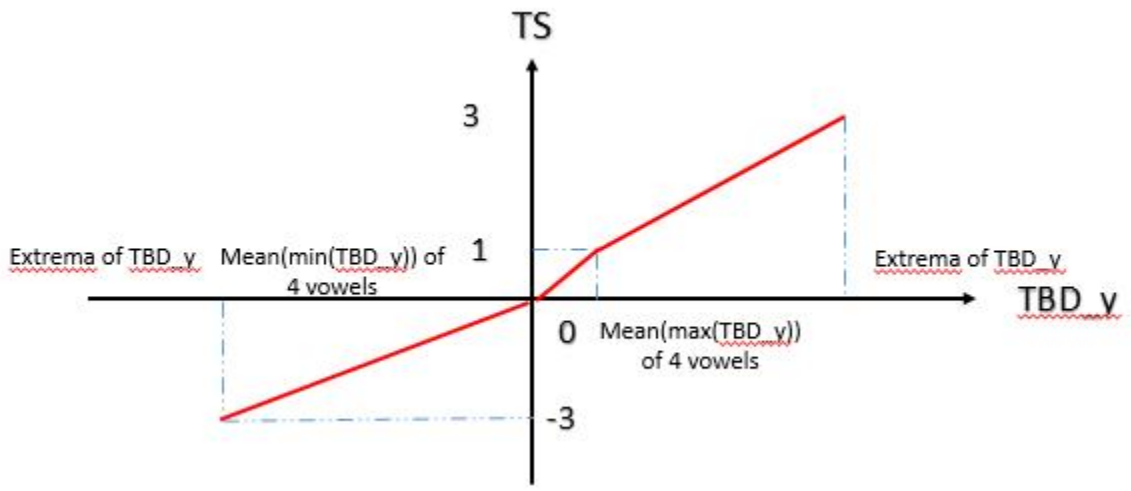

Figure 17: The 4 point piecewise mapping for the TS parameter

The linear mapping for the TA parameter is shown in Figure 18 with specific coordinates as follows:

- Extreme min value of TBD_y correspond to 3; (TBD_y _ext_min,3)
- Mean minimum value of vowel records correspond to 2 (TBD_y _vowel_min, 2)
- Mean maximum value of vowel records correspond to -0.5 (TBD_y _vowel_max, -0.5)
- Extreme max value of TBD_y correspond to -3 (TBD_y _ext_max,-3)

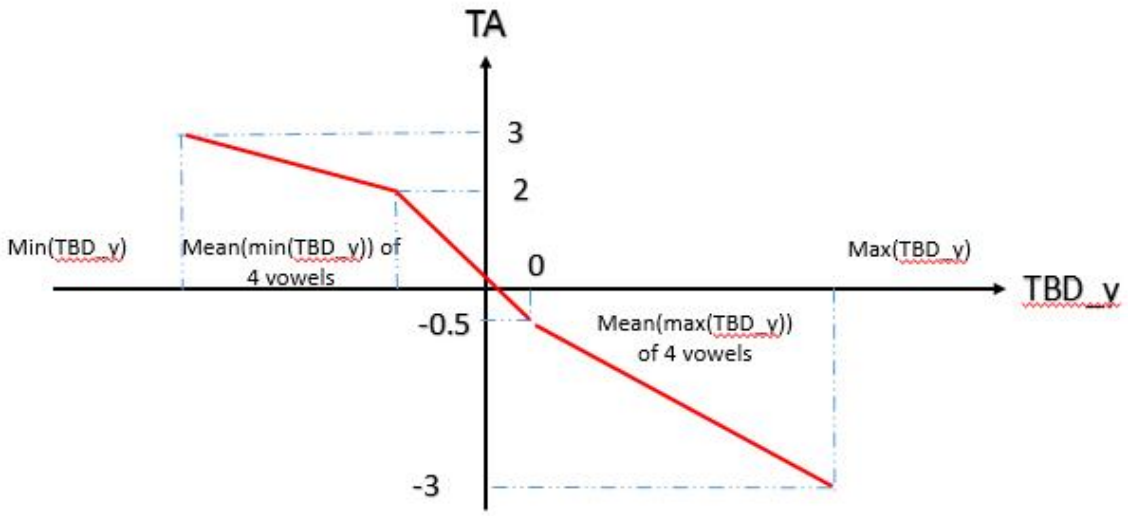

Figure 18: The 4 point piecewise mapping for the TA parameter

The linear mapping for the LA parameter is a two-point mapping, as shown in Figure 19 with specific coordinates as follows:

- Extreme min value of UL_LL correspond to -1.5; (UL_LL_min,-1.5)
- Extreme max value of UL_LL correspond to +3 (UL_LL_max,+3)

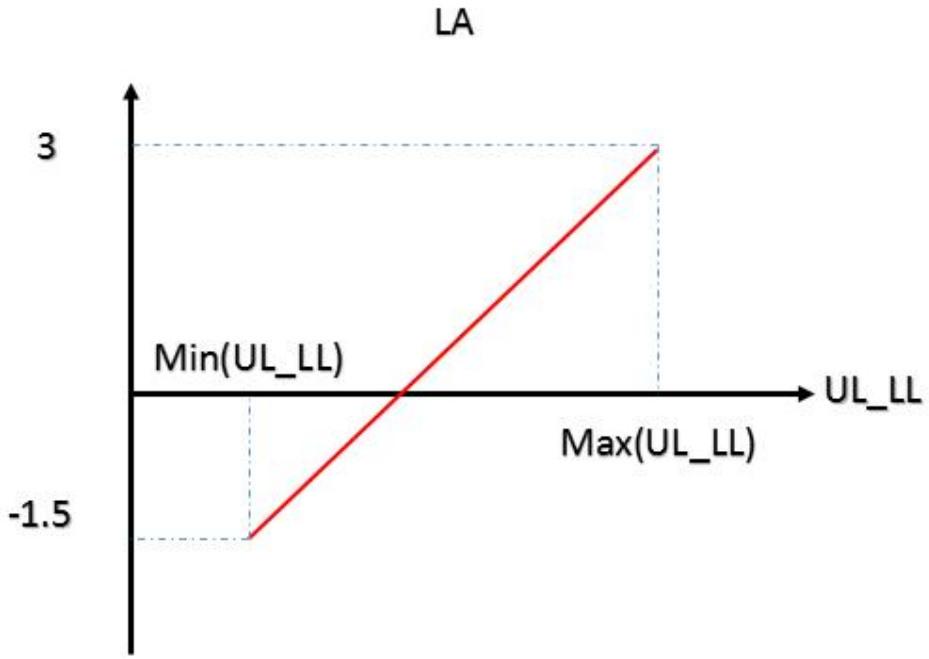

Figure 19: The linear mapping for the LA parameter

The linear mapping for the LP parameter is also a two-point mapping, as shown in Figure 20 with specific coordinates as follows:

- Extreme min value of LL_x correspond to -3; (LL _x_ext_min,-3)
- Extreme max value of LL_x correspond to +3 (LL_x_ext_max,+3)

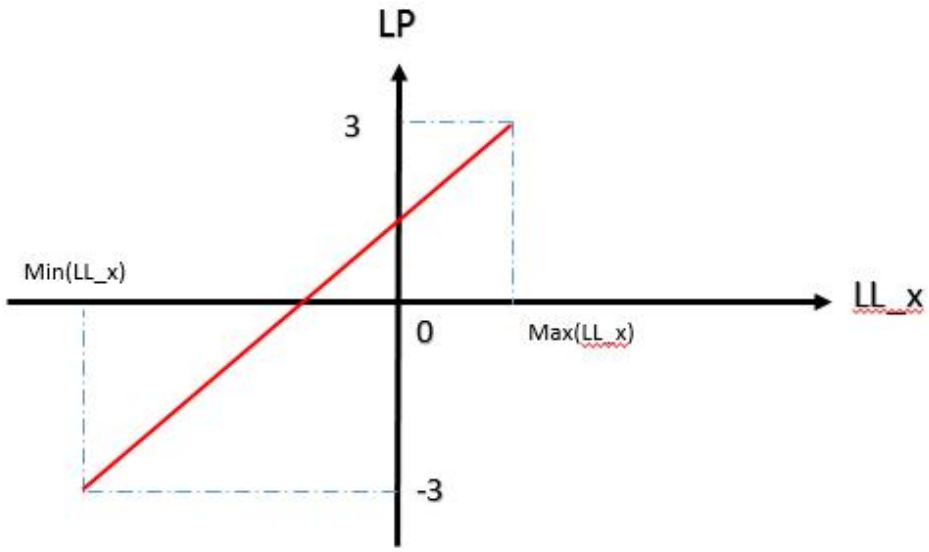

Figure 20: The linear mapping for the LP parameter

All of the above mappings are implemented on the kinematic data as direct linear interpolations on a sample by sample basis in real-time. For each synthesis parameter, the source kinematic measurement is used first to select the appropriate linear segment and then to interpolate the output synthesis parameter value.

2.4.2. 2-point piecewise linear mapping method

The 2-point piecewise linear mapping is similar to the 4-point method, except that it is implemented using only the extrema data for a subject. The two end-points are determined from the minima and maxima values of the extrema speech record Details of this mapping are as follows:

Application sensors: TB, TD, MI, UL and LL

Synthesis parameters controlled: LP, LA, TH, TE, TP, and JW

Synthesis parameters fixed settings:

- Larynx Height (LH): Fixed at 0
- Glottal Aperture (GA): Fixed at continuous voicing
- Fundamental Frequency (FX): Fixed at 0 (default value)
- VP Opening (NS): Fixed at VP Closed

The specific articulatory mappings:

- The average Y position of the MI sensor is used to directly calculate the value MI_y which maps onto the JW parameter.
- The average X positions of the TB and TD sensors is the value TBD_x which inversely maps onto the TP parameter
- The average Y positions of the TB and TD sensors is the value TBD_y which maps onto the TS parameter
- The average Y positions of the TB and TD sensors is the value TBD_y which inversely maps onto the TA parameter
- The Euclidean distance between the UL and LL sensors is the value UL_LL which maps onto the LA parameter
- The X position of the LL sensor is used to directly calculate the value LL_x which maps onto the LP parameter

The linear mapping for the JW parameter is shown in Figure 21.

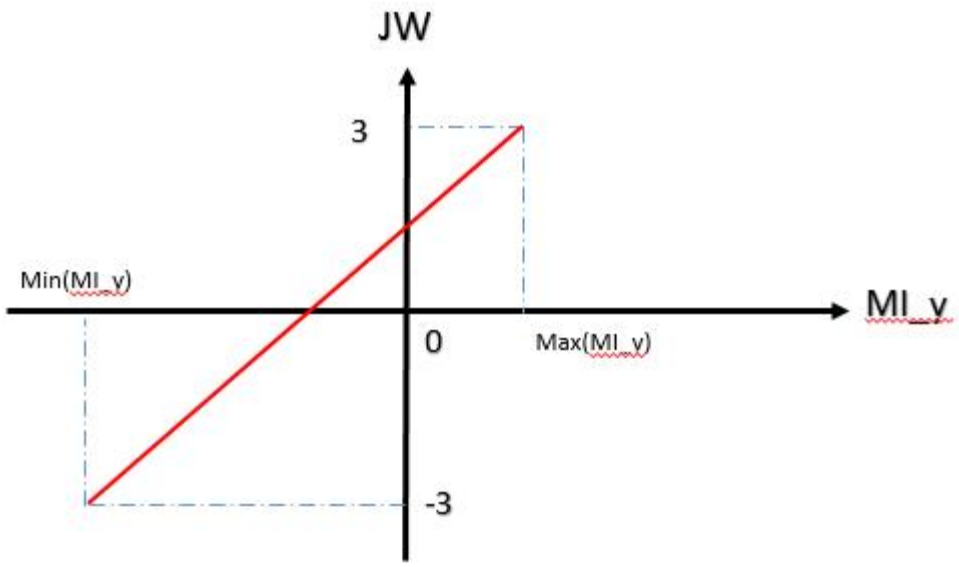

Figure 21: The 2 point piecewise mapping for the JW parameter

The linear mapping for the TP parameter is shown in Figure 22.

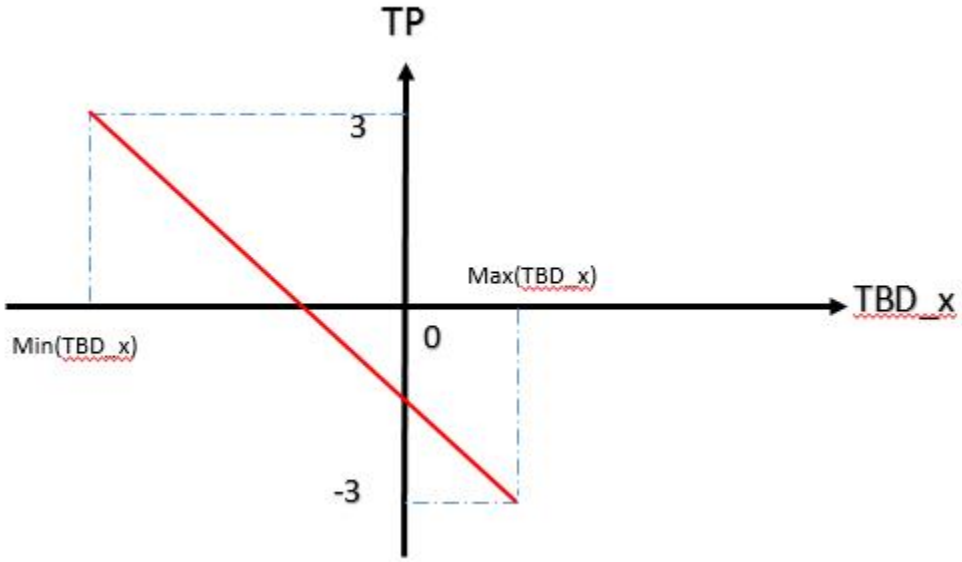

Figure 22: The 2 point piecewise mapping for the TP parameter

The linear mapping for the TS parameter is shown in Figure 23.

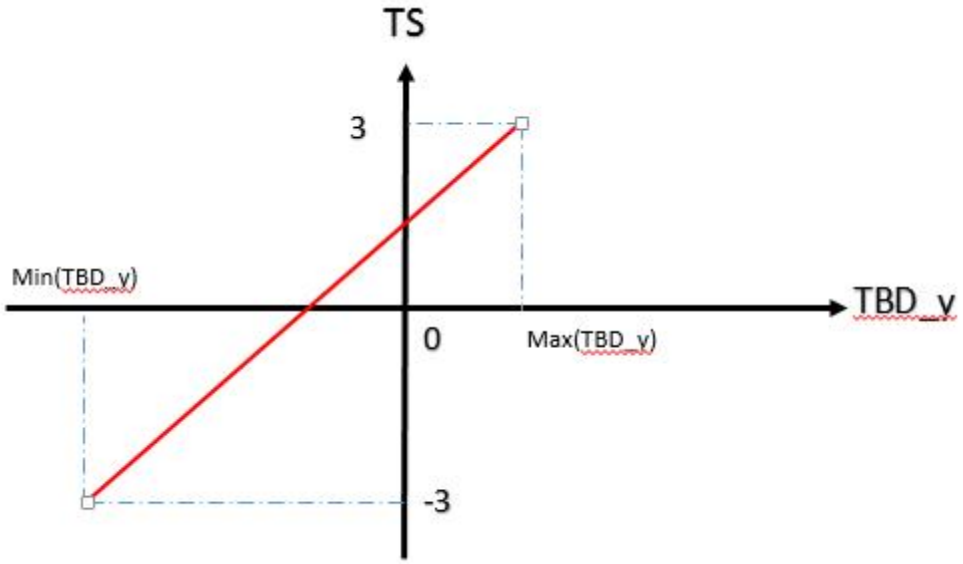

Figure 23: The 2 point piecewise mapping for the TS parameter

The linear mapping for the TA parameter is shown in *Figure 24*.

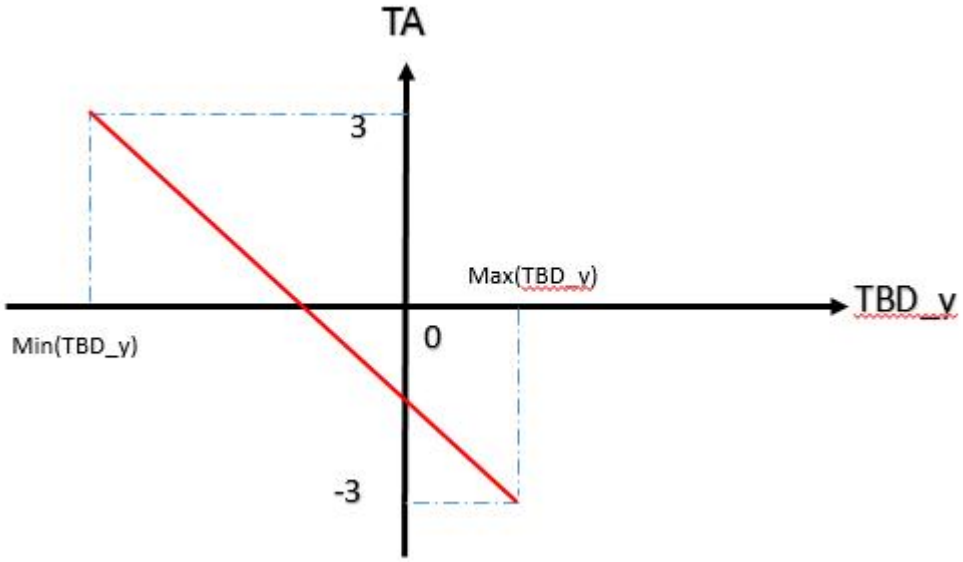

Figure 24: The 2 point piecewise mapping for the TA parameter

The linear mapping for the LA and LP parameter is the same as for the 4-point method, as shown previously in Figure 19 and Figure 20.

2.4.3. Quantile-based mapping method

In the quantile mapping approach, the mapping is based on the probability distribution of the selected kinematic data variable across a continuous speech record. The original intent of this mapping was to divide the synthesis parameter range (-3 to +3) into 0.1 increments, giving 61 individual parameter values. These 61 values were then to be mapped to a kinematic value based on evenly separated breakpoints, or quantiles, of the probability distribution, e.g. the first value is set to the 0 percentile (lower extrema) value, the second value to 100*(1/61) = the 1.64 percentile value, and so on.

The implementation methodology of this quantile approach in the RASS system inadvertently resulted in a set of kinematic data values different than the even 0.1 increments originally intended. In practice, the "quantile" command from the standard C library was applied to the integer synthesis parameter values {-3,-2,-1,0,1,2,3}, a set of N=7 data values. The algorithm assigns quantiles under 0.5/N to the minimum value, above (N-0.5)/N to the maximum value, and uses linear interpolation to determine quantile values between those. This resulted in the target {0, 1/61, 2/61, … 60/61, 61/61} quantiles being assigned to the values:

{-3, -3, -3, -3, -2.9167, -2.8000, -2.6833, …, 2.8000, 2.9167, 3, 3, 3, 3}

with a synthesis parameter interval of 0.1167.

The LA synthesis parameters is assigned values in the range -1.5 to +3, because the lower values are not used for normal voice synthesis. In this case the quantile

algorithm on the value set {-1.5, -1, -0.5, 0, .5, 1.0, 1.5, 2.0, 2.5, 3.0} with N=10 resulted in the same {0, 1/61, 2/61, … 60/61, 61/61} target quantiles being assigned the values:

{-1.5, -1.5, -1.5, -1.4167, -1.3333, …, 2.7500, 2.8333, 2.9167, 3, 3, 3, 3}

with a synthesis parameter interval of 0.0833.

Details of this method are as follows:

Application sensors: TB, TD, MI, UL and LL

Synthesis parameters controlled: LP, LA, TH, TE, TP, and JW

Synth parameters fixed settings:

- Larynx Height (LH): Fixed at 0
- Glottal Aperture (GA): Fixed at continuous voicing
- Fundamental Frequency (FX): Fixed at 0 (default value)
- VP Opening (NS): Fixed at VP Closed

The specific articulatory mappings:

- The average Y position of the MI sensor is used to directly calculate the value MI_y which maps onto the JW parameter, shown in Figure *25*.
- The average X positions of the TB and TD sensors is the value TBD_x which inversely maps onto the TP parameter, shown in Figure *26*
- The average Y positions of the TB and TD sensors is the value TBD_y which maps onto the TS parameter, shown in Figure *27*

- The average Y positions of the TB and TD sensors is the value TBD_y which inversely maps onto the TA parameter, shown in Figure *28*
- The Euclidean distance between the UL and LL sensors is the value UL_LL which maps onto the LA parameter, shown in Figure *29*
- The X position of the LL sensor is used to directly calculate the value LL_x which maps onto the LP parameter, shown in Figure *30*

Examples of the resulting mappings are shown for each synthesis parameter in Figure 25 to Figure 30 below. Values between the 61 identified points in each mapping are linearly interpolated between the two adjacent mapping values.

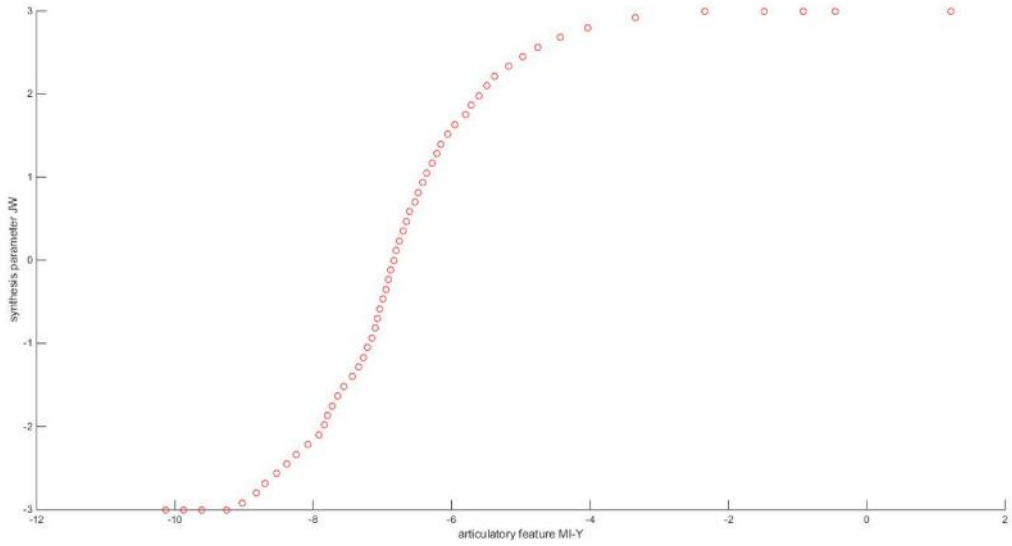

Figure 25: An example quantile mapping for the JW parameter

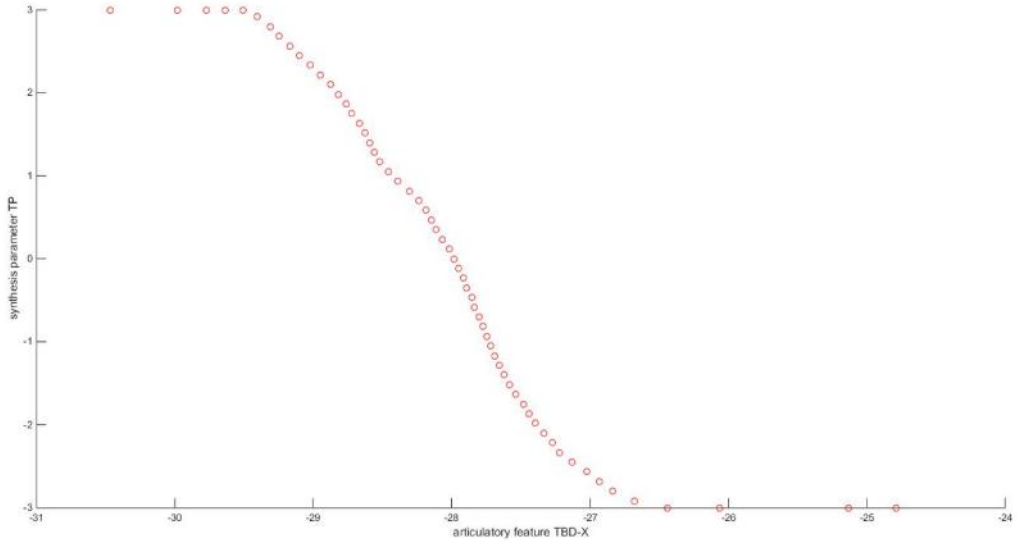

Figure 26: An example quantile mapping for the TP parameter

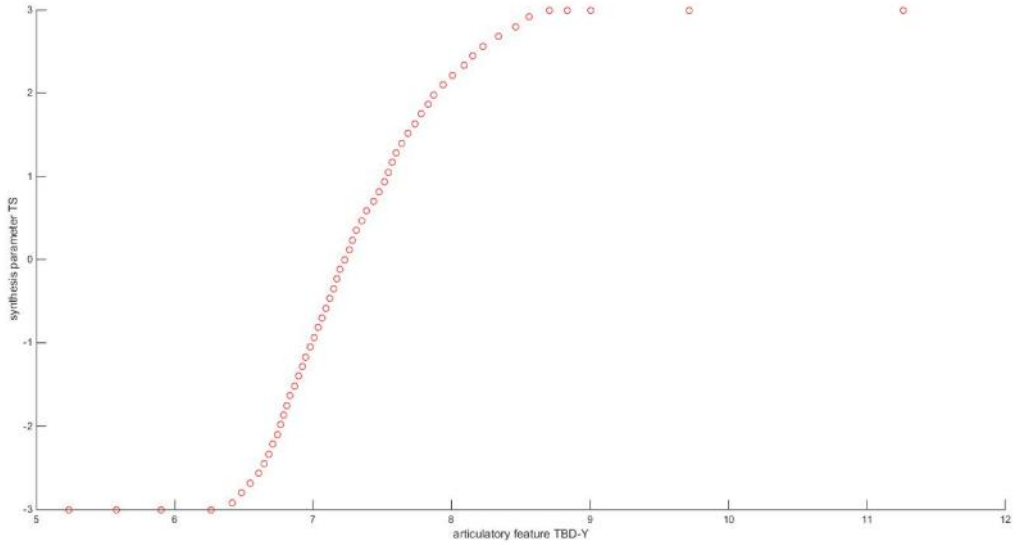

Figure 27: An example quantile mapping for the TS parameter

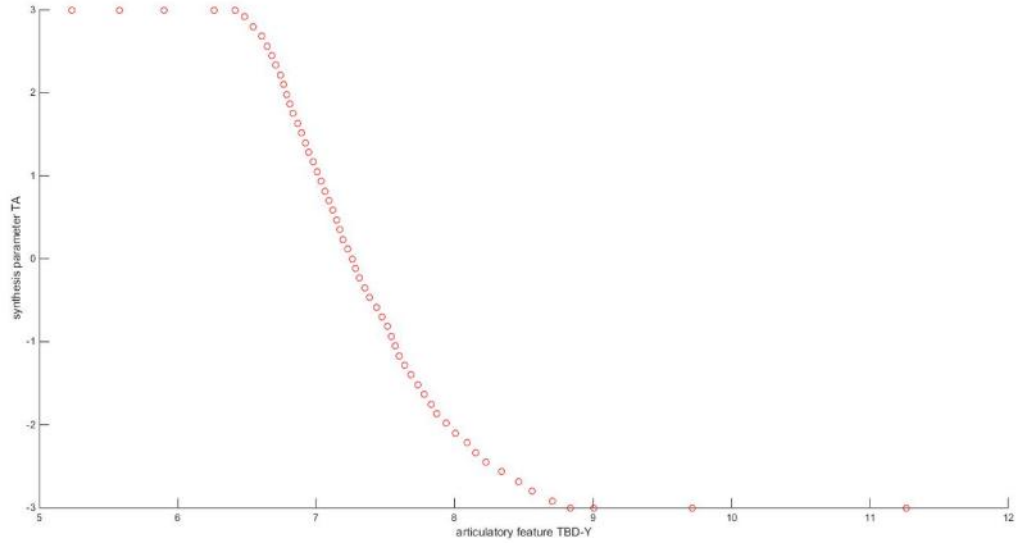

Figure 28: An example quantile mapping for the TA parameter

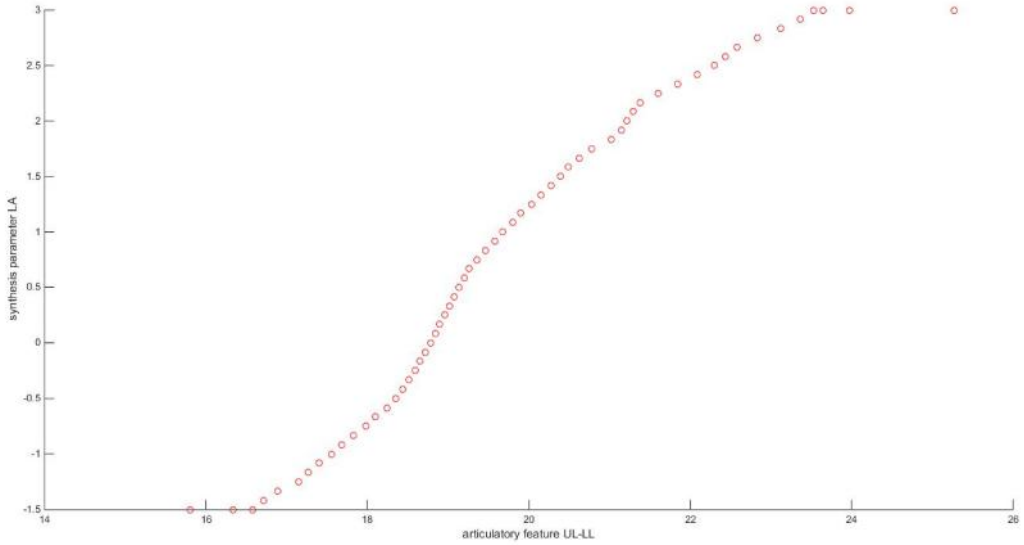

Figure 29: An example quantile mapping for the LA parameter

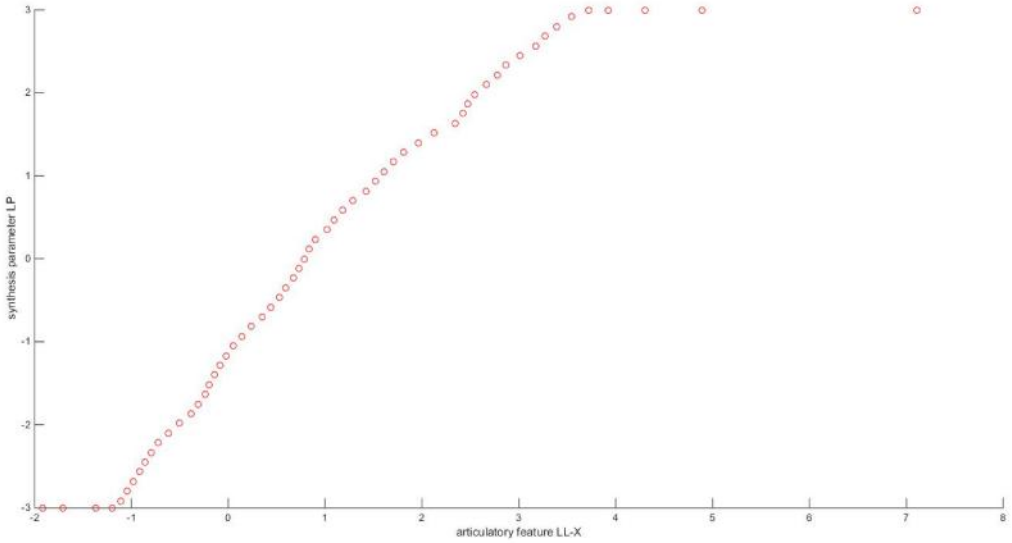

Figure 30: An example quantile mapping for the LP parameter

2.4.4. Discussion about previous mapping method

The simple connection between an individual kinematic position value and an individual synthesis parameter, implemented by all these prior methods, is a substantial simplification of the complex relationship between articulatory motion and vocal tract configuration. Because of the overly simplified model, the experimenter running the RASS system often needs to individually adjust the coordinates of these mapping methods to get adequate performance. Synthesis is typically limited to stationary vowel configurations, since a much more precise mapping that includes specific types of closure points would be necessary to adequately synthesize consonants.

2.5. Audapt System

In some RASS experiments, the resulting synthesized audio is further modified to change formant values before playing them back to the subject. The Audapt system is used to do this. Audapt is a speech manipulation software tool that both collects and manipulates speech files. It allows precise control over acoustics of a subject's speech in real time. Audapt returns the modified speech to the speaker through headphones. Figure 31 displays the Audapt GUI and related variables.

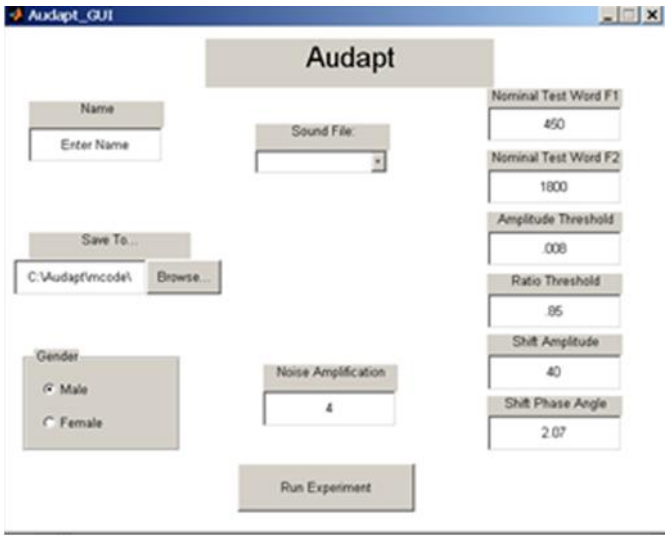

Figure 31: Audapt GUI

2.6. Summary

This chapter has introduced the NDI Wave (EMA) system, the RASS system, the VTDemo synthesizer, previous mapping methods, and the Aduapt system. The goal is to give a background of the entire system and compare as well as contrast the different methods so that the theory can be extended to a new more accurate mapping methods

CHAPTER 3 Palate Mesh creation

In this chapter, the fundamental concepts of three dimensional palate trace creation and related evaluation experiments will be introduced.

3.1. Introduction

One of the most important aspects of mapping from kinematic articulatory data to synthesis parameters is representing the kinematic data in a way that most accurately conveys the acoustic structure of the vocal tract. In order to do this, the kinematic data needs to relate as closely as possible to the vocal tract shape, specifically the cross-sectional area of the vocal tract. Sensor position information does not accomplish this directly, primarily because it does not capture the vocal tract boundaries, the most significant of which are the hard and soft palate of each subject.

In order to more accurately capture vocal tract configuration, we introduce a method for using the information from the palate trace to create more meaningful articulatory features for synthesis parameter calculation. The first step of this approach is the creation of an accurate palate representation from the palate trace record taken from each subject during the calibration stage of data collection. The method introduced in this chapter for capturing the palate shape of subjects is based on a modified Thin Plate Spline (TPS) [9] approach. The TPS approach is augmented using a grid-based method that keeps only the vertically highest data points in each grid, in order to reduce the possibility of outlier data caused by experimenter error during palate trace capturing.

For evaluation of the proposed approach, we use a data subject with no known outlier effects, and create outliers artificially to mimic the target scenario. These artificial data points simulate the behavior of a palate wand moving off of the palate for a short period of time. Results are evaluated using mean-squared error to the TPS-generated palate on the original data with no outliers.

3.2. Thin-plate spline method

The thin-plate spline algorithm [9] is an established method for generating a palate estimate for each subject. The basic idea of this approach is to start with a flat plane, and then warp this plane in a way that both fits the collected data and meets pre-specified smoothness constraints.

The thin-plate smoothing spline f is the unique minimizer of the weighted sum

$$\mathrm{pE}(f) + (1-p)R(f)$$

with E(f) the error measure

$$\mathrm{E}(f) = \sum_j \left| y(:,j) - f\big(x(:,j)\big) \right|^2$$

and R(f) the roughness measure

$$\mathrm{R}(f) = \int (|D_1 D_1 f|^2 + 2|D_1 D_2 f|^2 + |D_2 D_2 f|^2).$$

Here, the integral is taken over all of R^2, and D denotes the partial derivative of f with respect to its ith argument, hence the integrand involves second partial derivatives of f.

The smoothing parameter p is chosen so that (1-p)/p equals the average of the diagonal entries of the matrix A, with A + (1 − p)/p □ eye(n) the coefficient matrix of the linear system for the n coefficients of the smoothing spline to be determined. This choice of p is meant to ensure that we are in between the two extremes of interpolation (when p is close to 1 and the coefficient matrix is essentially A) and complete smoothing (when p is close to 0 and the coefficient matrix is essentially a multiple of the identity matrix). The smoothness factor is an important component [10]. According to prior results from the TPS surface reconstruction technique [10], we use the preferred smoothing parameter P = 0.95.

This approach works well for accurate palate trace data where the sensor has maintained consistent contact with the palate surface during the entire recording. Figure 32 shows the TPS-derived palate mesh for subject 02 of the EMA-MAE dataset, which has a thoroughly covered palate trace record with no identifiable outliers.

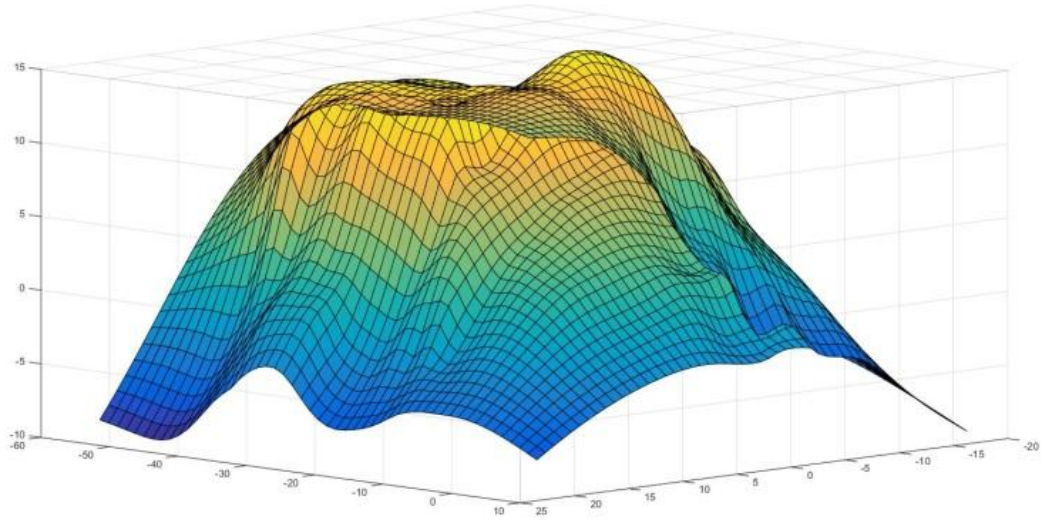

Figure 32: Palate mesh for EMA-MAE subject 02

However, this method has difficulty when there are problems with the initial palate data. During EMA palate trace collection, the experimenter physically moves the plate trace wand in a pattern across the subject's palate. It is very common to have points in the record where the wand moves off the palate, for example when changing direction of motion or when the sensor tip comes into contact with a small bump. To illustrate this, Figure 33 shows the TPS-derived palate mesh for subject 18 of the EMA-MAE dataset, which has several outlier segments in the palate record. The outliers are noted in the figure with a red circle.

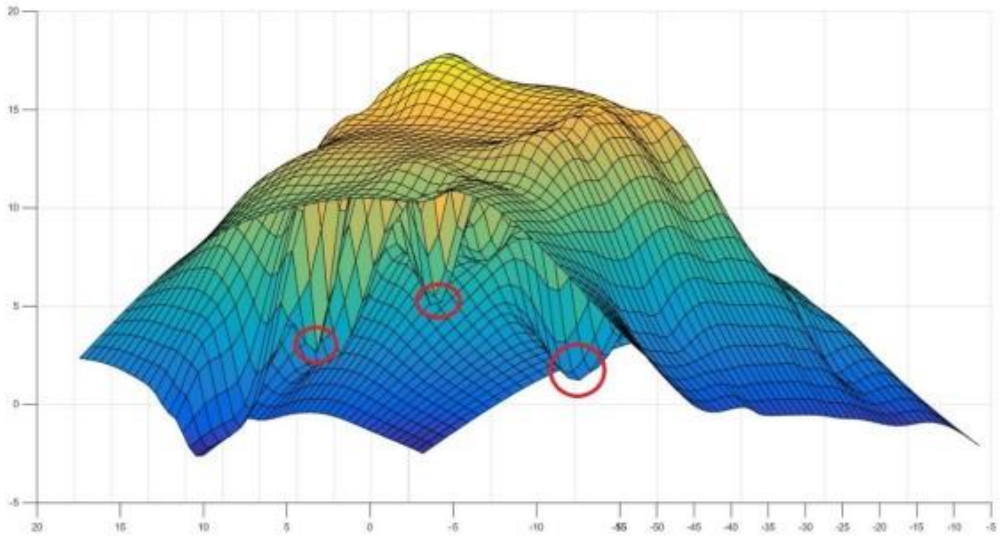

Figure 33: Palate mesh for subject18 with outliers marked by red circle

3.3. Proposed methods

3.3.1. Convex hull concept

The TPS method is a least-squares approach, which implicitly assumes a symmetric two-sided error pattern. However, because of the nature of the physical palate boundary, errors related to wand placement are by definition one-sided. An alternative surface fit that incorporates unidirectional error would be preferable. One such model would be based on a convex hull approach. The convex hull is a fundamental construction for mathematics and computational geometry [11], representing the outer boundary of a set of elements. Given a sufficient coverage of palate trace data, a convex hull over the data would represent an upper bound that matches the true palate surface. The intent

of this idea is to use the convexity constraint in a way that removes the outlier points discussed above from consideration.

There are algorithms, such as Matlab's Quickhull [12] implementation, which can return the convex hull of a set of points. However, this cannot be directly used for our palate application, because it returns the entire three-dimensional convex bound, including both upper and lower boundary surfaces. In addition, there are some implicit problems with the convexity constraint, because real palates may also include some concave regions that would be lost in this approach.

Alternatively, in order to capture the general idea of the convex hull as a method for removing outliers and off-palate data points, we instead use an approach that focuses on keeping the locally uppermost points, which will implicitly eliminate outliers. To do this, we use an implementation in which the data is first divided into grids and then the highest vertically-valued points in each grid are retained, prior to implementing the TPS fit.

3.3.2. Gridded convex hull method

Based on the convexity principle, we implement a new gridded convex hull method, which is a combination of a partitioned convex hull approach and a thin-plate spline for smoothing. The basic idea of this method is that we separate the whole palate trace into an n by n grid, and within each grid region we select a fixed percentage of the highest vertically valued points. We then remove the other points and then re-create the palate using the TPS method. The underlying idea is that these outliers caused by experimenter

error tend to happen in short segments within a small horizontal region, and are all of substantially lower vertical value, so they can be identified within a region and removed using a simple percentile threshold.

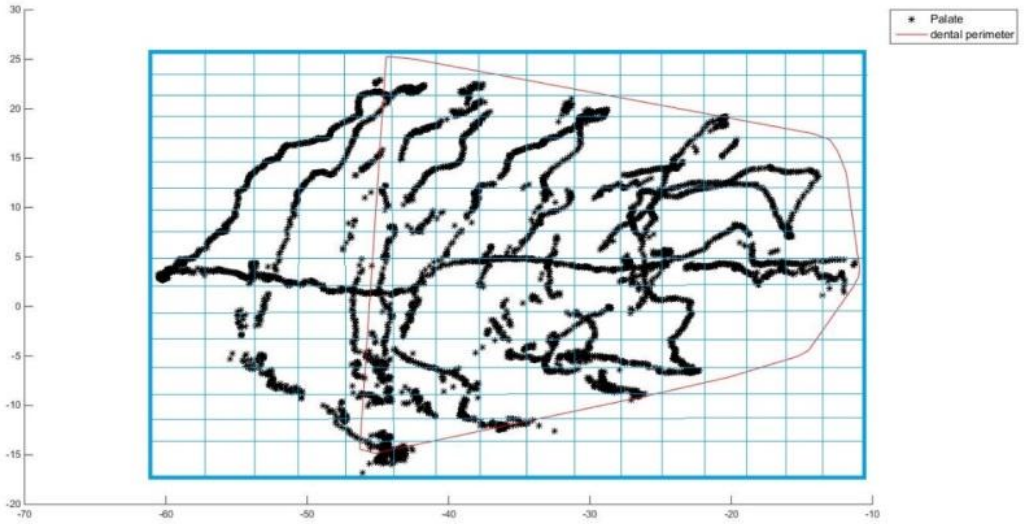

Figure 34: Gridded convex hull

The steps of the gridded convex hull method are as follows:

1. Identify the maximum and minimum values of the horizontal plane (X and Z axes) and calculate the step size according to the desired grid resolution to cover the palate trace. This was empirically varied to include 10x10, 20x20, 30x30, or 40x40 size grids.
2. Within each grid, keep a fixed percentage of the points with maximum vertical height (Y value). This was empirically varied to increment from 10 to 90 percent, in 10 percent increments.

3. Using only these uppermost points within each grid, implement the thin-plate spline algorithm.

3.3.3. Example results of gridded hull method

To process the palate trace recording files, the first step is to extract the entire palate sensor record, and then identify the start and end time of each section of the record, including the outer dental boundary, the inner dental boundary, and the main palate trace. The dynamic range of the palate was determined using dental boundaries, since in some cases the palate trace does not adequately cover the posterior portion of the palate region sufficiently.

The grid size and percentage of kept vertices in each grid was empirically varied. Grids included 1X1, 5x5, 10x10, 20x20, 30x30, and 40x40, and percentages of kept points ranged from 5% to 90 percent in 10 percent increments.

Examples of typical results are shown from Figure 35 to Figure 40. These figures illustrates the results with different grid size and percentage of kept points with the original palate trace record for comparison: Figure 35 shows the result of subject 18 after implementing a 10x10 gridded Convex Hull with 90% kept points, Figure 36 shows the result of subject 18 with a 20×20 grid with 90% kept points, Figure 37 shows the result of subject 18 with a 40×40 grid with 90% kept points, Figure 38 shows the result of subject 18 with a 10×10 grid with 95 percent kept points, Figure 39 shows the result of subject 18 with a 10×10 grid 85 percent kept points, and Figure 40 shows the result of subject 18 with a 10×10 grid with 10 percent kept points. According to these figures,

we can observe that different grid size and percentages lead to different shapes of palate trace.

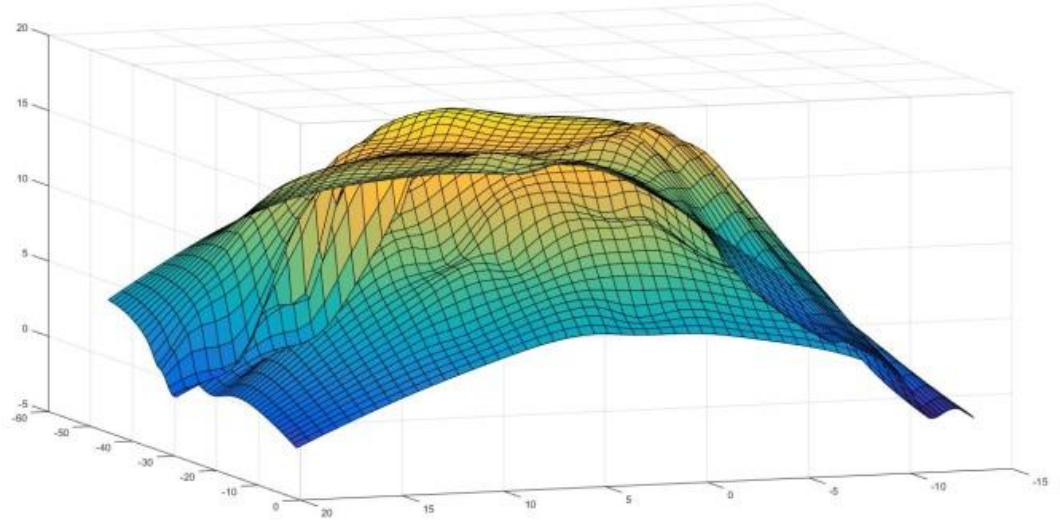

Figure 35: Palate mesh of subject 18 with grid size at 10x10 and 10% kept points

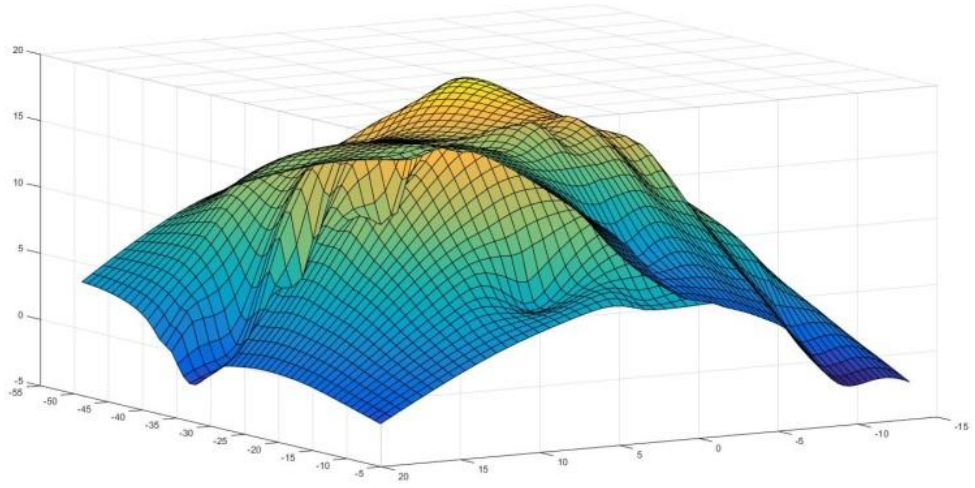

Figure 36: Palate mesh of subject 18 with grid size at 20x20 and 10% kept points

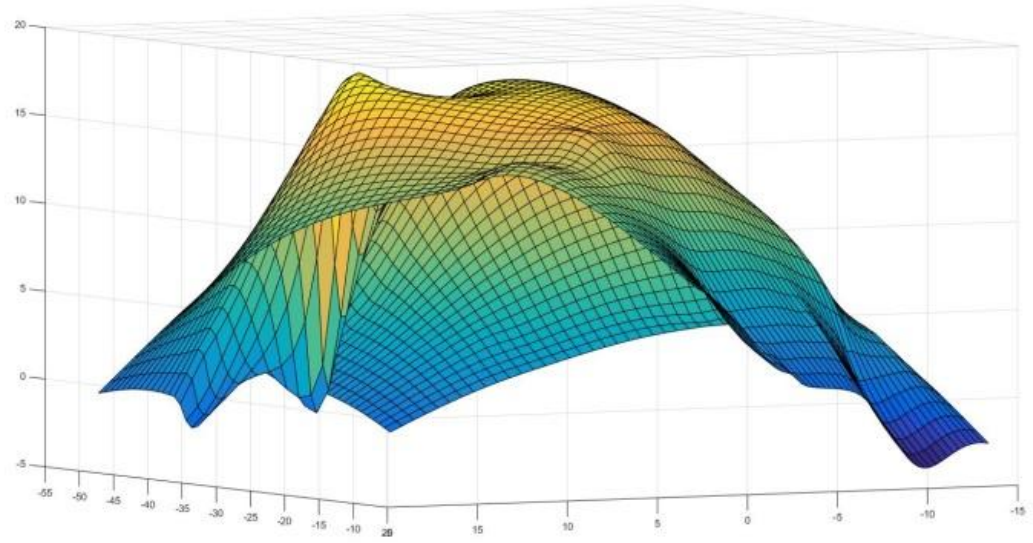

Figure 37: Palate mesh of subject 18 with grid size at 40x40 and 10% kept points

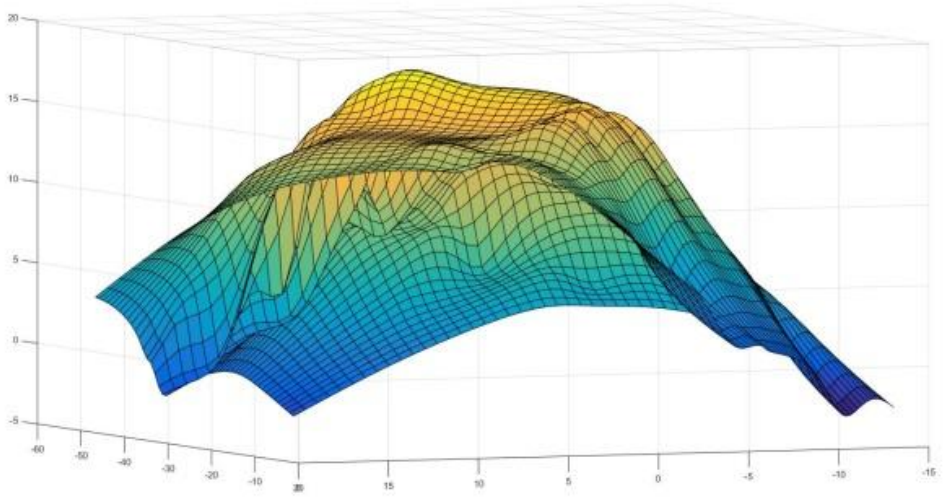

Figure 38: Palate mesh of subject 18 with grid size at 10x10 and 5% kept points

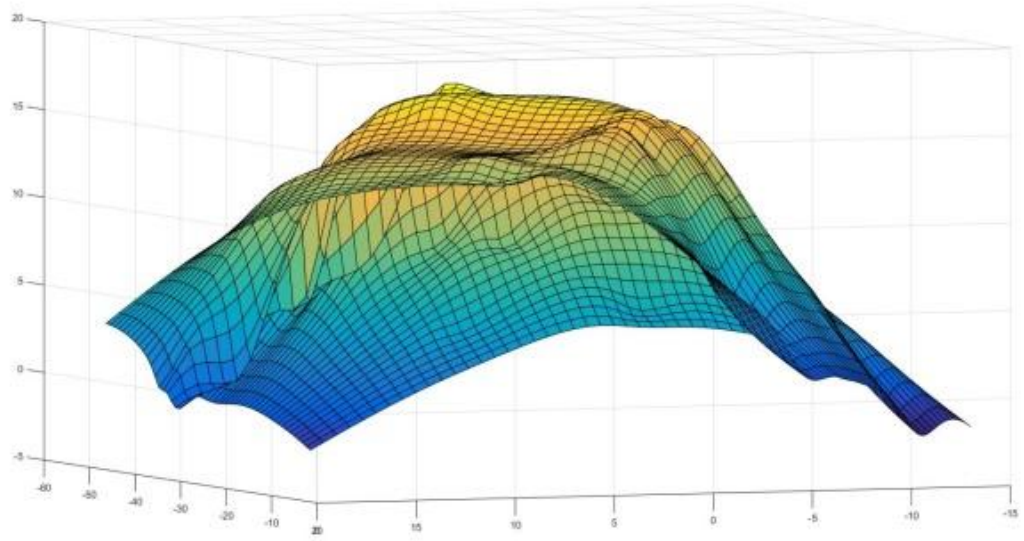

Figure 39: Palate mesh of subject 18 with grid size at 10x10 and 85% kept points

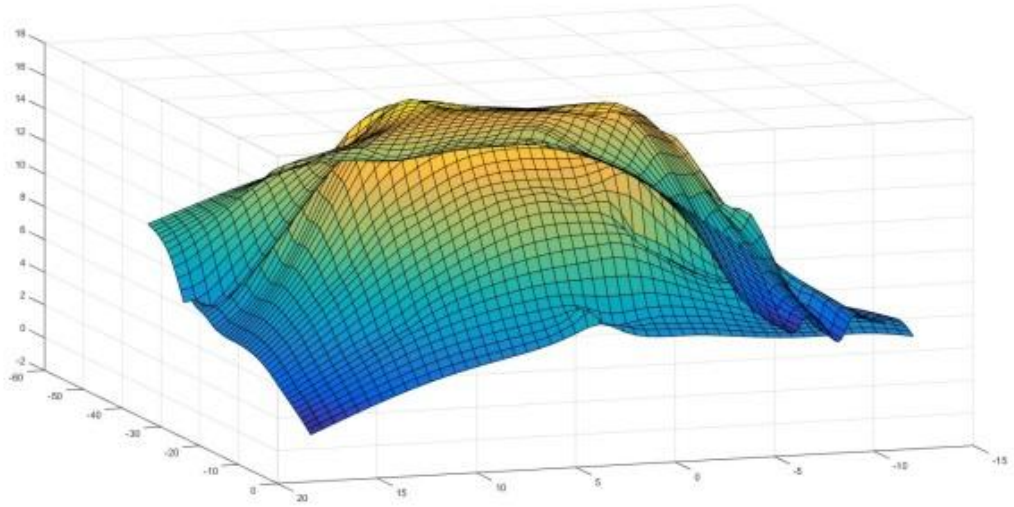

Figure 40: Palate mesh of subject 18 with grid size at 10x10 and 90% kept points

3.3. Evaluation methodology

For evaluating the gridded convex hull method, we need to compare the final palate mesh against a known correct palate mesh. This is problematic since none of the subjects which have outlier problems in the palate records have known correct palate meshes. To address this, we choose to create an artificial evaluation set using a subject with good palate data and no outliers, adding outliers that represent the kinds of errors we typically see. The artificial outliers are added using a linear interpolation method to simulate the wand being away from the subjects' palate during recording. To achieve this, we randomly select one point in the area of the palate trace, and insert an outlier segment at that point, with a height and time length randomly selected from within a uniform distribution determined empirically through analysis of the outliers in the data set. We then implement the percentile gridded convex hull method on the new palate trace data with the artificial outlier to verify the effectiveness and feasibility of the new method.

Specifically, an outlier segment is created by choosing a random downward angle, a random distance distributed uniformly between 6 and 10mm, and a random time period chosen uniformly between 100 and 500ms, and adding a straight line corresponding to these values. This gives roughly 40 to 200 points of simulated outlier values at the 400Hz kinematic sampling rate.

We created two simulated evaluation conditions, one created by adding a single outlier and one created by adding multiple random outliers, where the number of outliers was

chosen from 2 to 5. We then implemented the gridded convex hull method, and calculated the corresponding mean squared error between the resulting palate mesh and the baseline. For the one outlier case, we implemented 45 times in total, using a grid size of 1x1, 5x5, 10x10, 20x20, 30x30 and 40x40 and percentage of kept points ranging from 10% to 90% in 10% increments. For the random number of outliers, we used the same number of experimental implementations was 36 for grid size from 1x1, 5x5, 10x10 20x20,30x30 and 40x40, percentage rate from 5% to 90%, but for each one of these we executed the configuration 220 times, with a different chosen number (from 2 to 5) of outliers in each case.

Results are shown below. Figure 41 to Figure 43 are examples of the single outlier case and Figure 44 and Figure 45 are examples of the multiple outlier case.

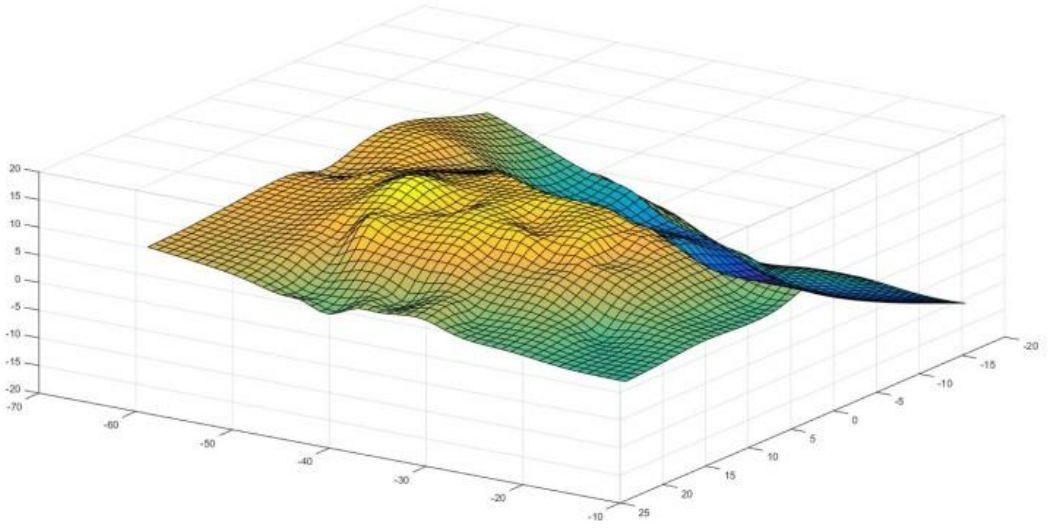

Figure 41: Original palate mesh of subject 1

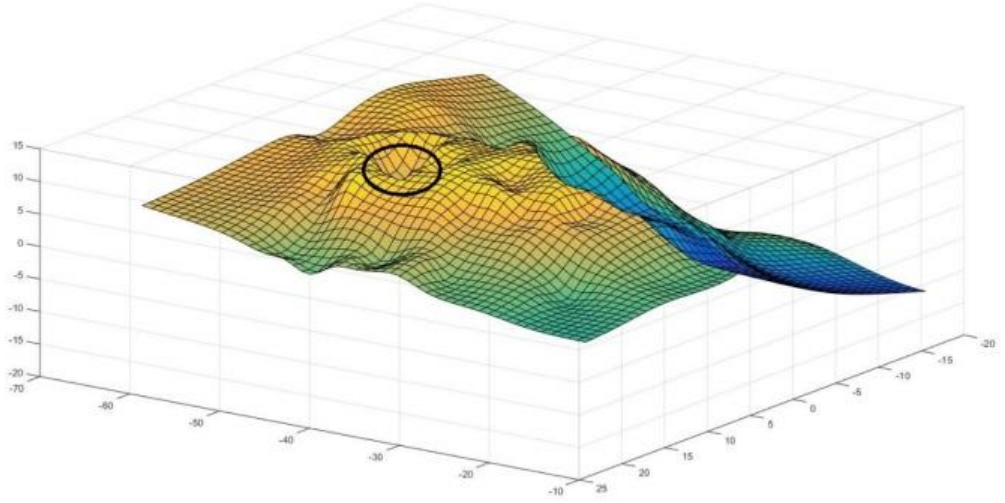

Figure 42: Palate of subject 1 with an artificial outlier

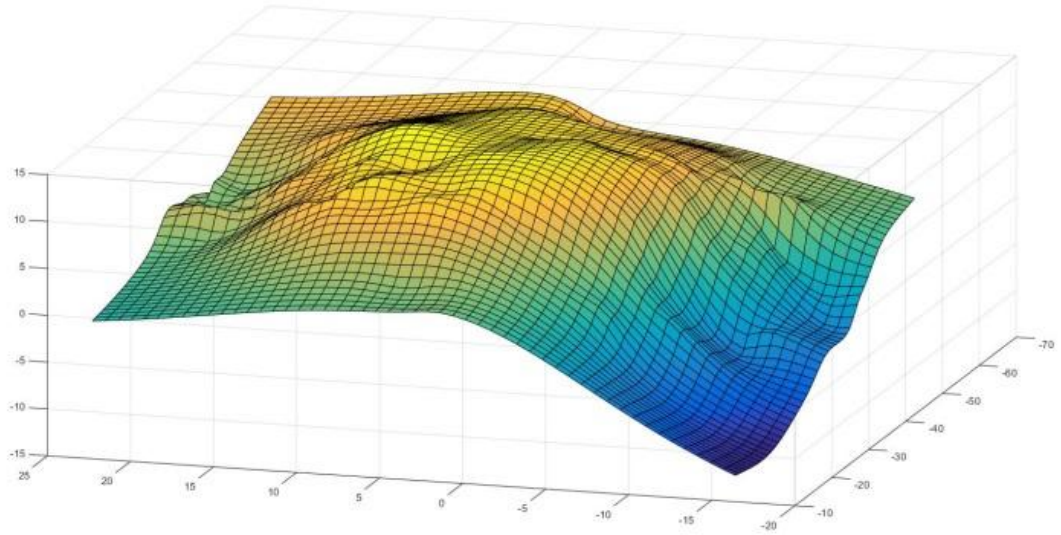

Figure 43: Palate mesh of subject 1 after using the gridded convex hull method with 10x10 with 10% kept points

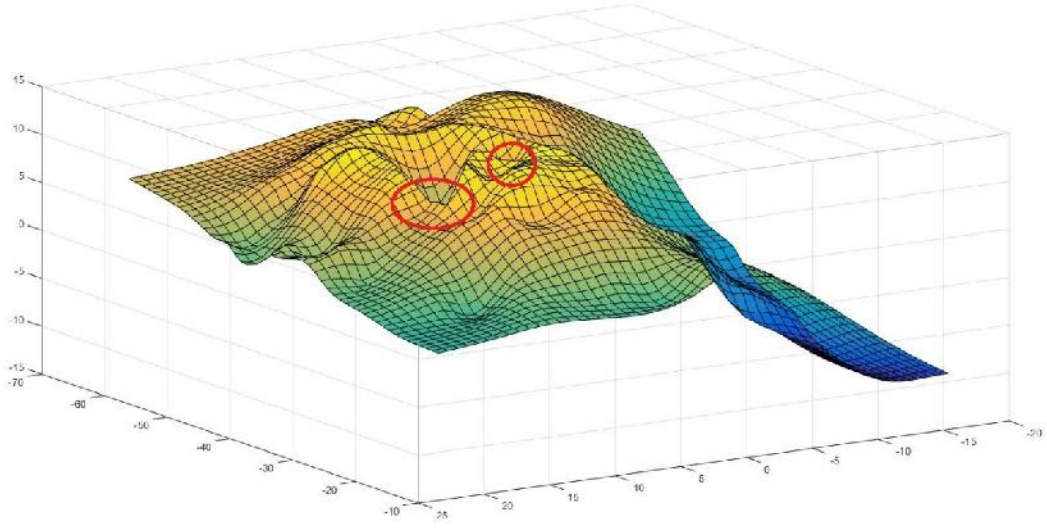

Figure 44: Palate of subject 1 with random number of artificial outliers

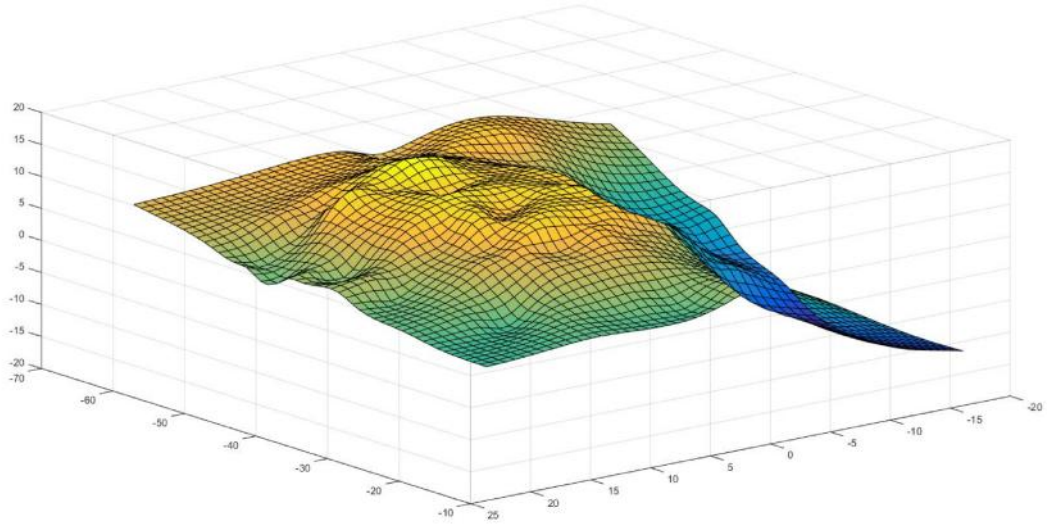

Figure 45: Palate mesh of subject 1 with random number of artificial outliers after using the gridded convex hull method with 10x10 with 10% kept points

From Figure 41 to Figure 45 illustrate how the gridded convex hull method is able to effectively remove this type of outlier.

3.4. Root mean square error versus grid size and percentage of kept vertices

For quantification of our method's accuracy, we calculate the root mean square error between the baseline palate trace and the processed palate trace after adding artificial outliers as described above and using the gridded convex hull method.

$$\text{RMSD} = \sqrt{\frac{\sum_{t=1}^{n}(\hat{y}_t - y)^2}{n}}$$

The following table shows the RMS error versus grid sizes and percentage of data kept in each grid. Table 2 shows the example of subject 01. Table 3 shows the average RMS value of subject 01 with random outliers.

Table 2： RMS data for subject01 with one outlier

	Grid size					
percentage of kept points	**1X1**	**5X5**	**10X10**	**20X20**	**30X30**	**40X40**
5%	0.3057	0.2885	0.1435	0.3094	0.5450	0.6693
10%	0.3382	0.3235	**0.1217**	0.6503	0.7482	1.9063
20%	0.5082	0.2819	0.3127	0.6350	0.7670	1.9453
30%	0.5981	0.4577	0.2969	0.6266	0.7803	2.0341
40%	0.5874	0.4774	0.3256	0.6253	0.7914	2.0471
50%	0.9033	0.5491	0.3523	0.6229	0.8047	2.0406
60%	1.6588	0.7755	0.4100	0.6227	0.8123	2.0276
70%	2.3023	0.7218	0.4559	0.6235	0.8214	2.0160
80%	3.1113	0.7582	0.4902	0.6206	0.8301	2.0077
90%	4.7766	1.1501	0.5382	0.6197	0.8389	2.0092

Table 3：Average value of RMS data for subject01 with randomly selected number of outliers

	Grid size					
percentage of kept points	**1X1**	**5X5**	**10X10**	**20X20**	**30X30**	**40X40**
5%	0.2151	0.1779	0.1778	0.3011	0.5352	0.6693
10%	0.2537	0.2377	**0.1576**	0.2569	0.5181	0.6359
20%	0.5083	0.3105	0.1903	0.3193	0.7302	0.7030
30%	0.5551	0.3515	0.2846	0.3368	0.5622	0.6018
40%	0.5893	0.5121	0.3147	0.3648	0.5949	1.0542
50%	0.8607	0.5253	0.3950	0.4619	0.5965	0.7180
60%	1.6262	0.5821	0.4748	0.5087	0.5411	0.7309
70%	2.2959	0.6761	0.5791	0.5920	0.6754	0.9167
80%	2.9255	0.7489	0.6854	0.6940	0.7643	0.9860
90%	4.7761	1.0703	1.0344	0.8335	0.7632	1.1378

From the table above, the combination of 10 by 10 grid with 10 percent of points kept in each grid gives the best empirical results for recovering the subject's correct palate shape.

3.5. Summary

This chapter has introduced robust palate mapping methods used to accurately estimate each subject's palate mesh data, as well as experimental data supporting the parameters to be used in implementation.

Chapter 4 Kinematic to Synthesis Parameter Mapping

In this chapter, a new mapping method will be introduced, based on a least-squares linear mapping from articulatory features to articulatory synthesis parameters. The selection and calculation of the articulatory features is described, followed by the details of the linear mapping algorithm, and evaluation experiments comparing the mean squared error of the new mapping method to previous approaches.

4.1. Overview

As described in Section 2.1, the fundamental goal of mapping kinematic data to synthesis parameters is to represent the precise relationship between kinematics and acoustics in order to enable accurate acoustic synthesis. In order to accomplish this task, we first address the issue of identifying the most acoustically relevant articulatory features from the kinematic data. Based on the three-dimensional palate trace in relation to the sensor data and the principle of pronunciation with vowel and consonant, articulatory features can be calculated that relate directly to the configuration and cross-sectional area of the vocal tract. These articulatory features can then be used as input variables to a matrix-based linear mapping, trained using vocalizations for which the correct synthesis parameters are known. To learn the mapping, the well-known pseudo-inverse method and target synthesis parameters from phoneme identification and extracted from formant space are used. Using data from 5 native speakers (3 male, 2 female) in the EMA-MAE corpus, the mapping is then

compared to previous mappings using both formant distortion and PESQ, based on a mean-square error metric, described in Section 4.5.

4.2. Articulatory features

The selection of articulatory features is based on the goal of representing physical characteristics that correlate with acoustics, such as vocal tract structure and cross-sectional area. Given the palate mesh and the placement of the kinematic sensors, the most relevant features are those which directly represent the forward position of the tongue and the height of the vocal tract opening at the sensor locations. Based on this idea, the following 9 articulatory features are selected:

1. vertical distance from tongue blade sensor to palate; (AF1)
2. horizontal distance from tongue blade sensor to upper incisor: (AF 2)
3. lateral distance between tongue blade and tongue lateral sensors: (AF3)
4. Euclidean distances between all 3 pairs of tongue sensors (TB, TD and TL): (AF4)
5. vertical distance between upper and lower lips (lip opening): (AF 5)
6. lateral distance from upper lip to lateral lip sensor (lip width): (AF 6)
7. vertical distance from lower incisor (jaw) sensor to palate: (AF 7)

To provide data for training the articulatory-to-synthesis mapping, we selected 16 target phonemes including 8 vowels and 10 consonants. Phoneme boundary information for target vowels and consonants was provided through manual segmentation from trained students in the Marquette Speech and Swallowing Laboratory. Then, for each target vowel and consonant, the average value of articulatory features was computed, and a

corresponding set of synthesis parameters was selected using two different approaches, as described in Section 4.3.

The mechanism for identifying boundary points in the target phonemes varied depending on whether it was a vowel or a consonant, as described in more detail below.

4.2.1. Feature computation window for vowels

There are 8 target vowel phonemes used in these experiments: “i”, “I”,”e”,”æ”,”u”,” ʊ”, “o” and “a” (IPA notation). These are labeled as phoneme IDs from 1 through 8, respectively. Articulatory features from these vowels were calculated from a frame of speech centered at the labeled midpoint of the vowel, with ±10ms on either side. Figure 46 illustrates this phoneme identification window for vowels. In the figure, the range between two blue lines is original vowel area, the range between two red lines is phoneme identification window. A single articulatory feature value is computed for each vowel as the average value of the articulatory features in the phoneme identification window.

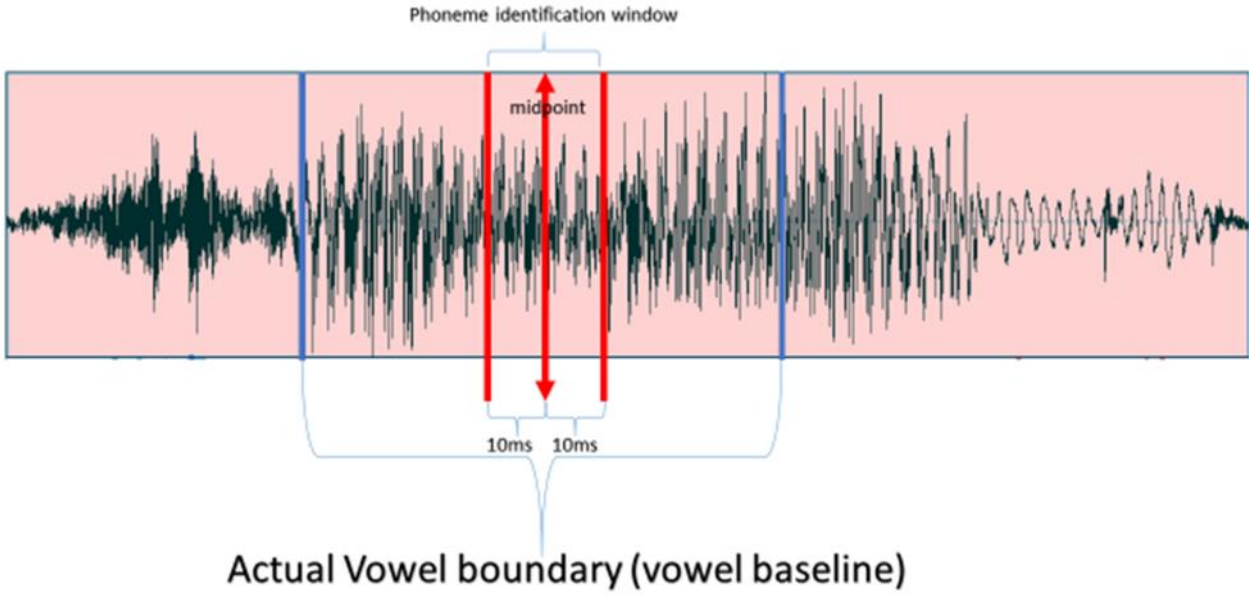

Figure 46: Phoneme identification window for vowels. The area between the two blue lines is the original vowel and the area between the two red lines is the phoneme identification window averaged to determine articulatory values.

4.2.2. Feature computation window for consonants

There are 10 target consonants used in these experiments, which are further divided into five groups according to the place and manner of articulation: 'b' and 'p' bilabial stops are group 1 (type ID 11) , 'd' and 't' alveolar stops are group 2 (type ID 12) , 'g' and 'k' velar stops are group 3 (type ID 13), 'f' and 'v' labiodental fricatives are group 4 (type ID 14), and 's' and 'z' coronal sibilants are group 5 (type ID 15). To identify the boundary regions for these consonants, we first chose a phoneme identification window 200 ms before the start of the following vowel. After this, we identify a specific closure point in this phoneme identification window. For group 1, we chose the identification point where the minimum lip distance occurs as the spot for articulatory feature calculation,

with an identification window for articulatory feature calculation centered at this spot ±10ms. For groups 2 and 5, we select the identification point where the minimum Y value of the TB sensor occurs, with an identification window for articulatory feature calculation centered at this spot ±10ms. For group 3, we choose the identification point with minimum Y value of the TD sensor, with an identification window for articulatory feature calculation centered at this spot ±10ms. For group 4, we choose the identification point which has minimum lip distance, with identification window for articulatory feature calculation centered at this spot ±10ms.

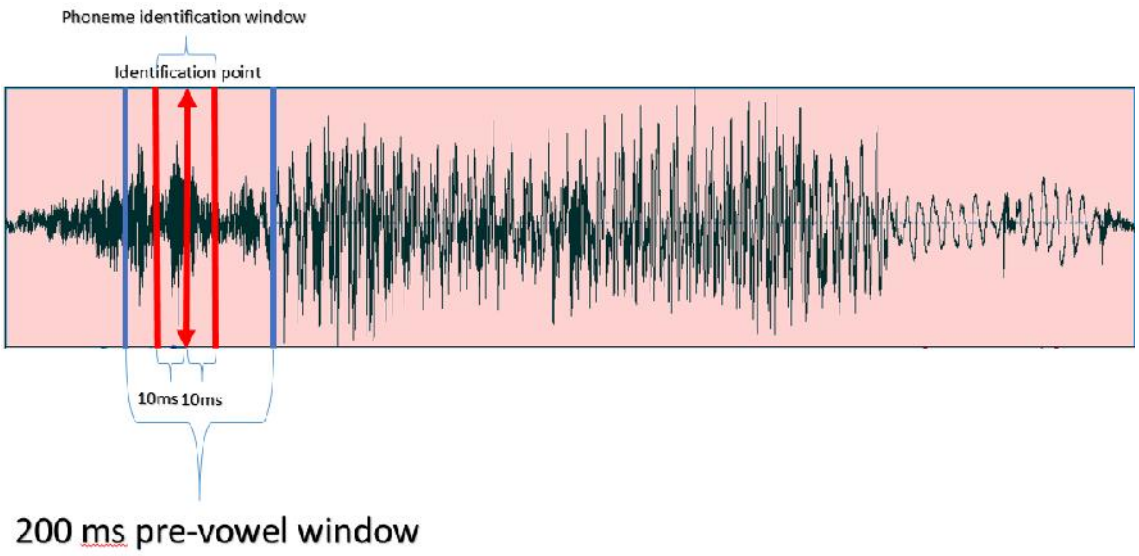

Figure 47 illustrates this phoneme identification window for consonants. In the figure, the range between two blue lines is 200ms consonant area before vowel area, the range between two red lines is phoneme identification window. A single articulatory feature value for each consonant is calculated as the average value of articulatory features in the phoneme identification window.

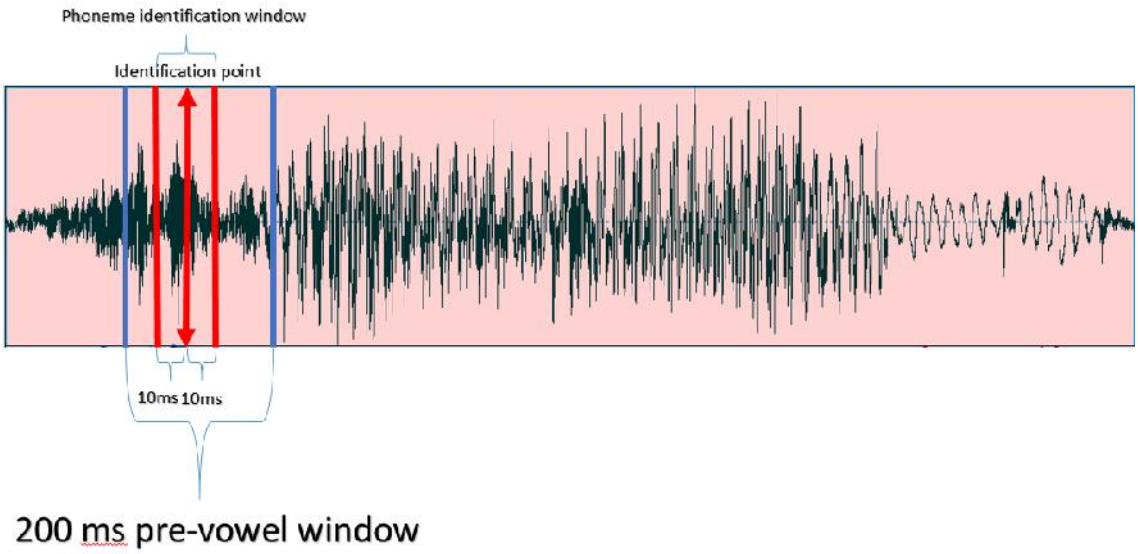

Figure 47: Phoneme identification window for consonants, the area between the two blue lines is the 200ms preceding vowel onset, while the area between the two red lines is the phoneme identification window averaged to determine articulatory values.

4.3. Target synthesis parameters

It is important to identify accurate synthesis parameters corresponding to the consonant and vowel data, in order to create an accurate mapping system. In these experiments we have implemented two different approaches to selecting synthesis parameters for training:

1. (All phonemes) Fixed synthesis parameters based on target phoneme type. In this approach, pre-determined synthesis parameters are selected from a table based on the corresponding phoneme type.

2. (Vowels only) Synthesis parameters determined for each example based on formant values. In this approach, acoustic F1 and F2 formant values are estimated from the vowel examples, and then a lookup table is used to identify synthesis parameters that most closely correspond to those formant values.

Each of these approaches is described in more detail below.

4.3.1. Target synthesis parameters based on phoneme type

In this approach, the target synthesis parameters are based entirely on the phoneme identity, from a table of prototypical values based on Maeda's original work [14]. These represent synthesis parameters, which will generate these specific phonemes, including the stationary configuration for vowels and the closure points for consonants, based on the corresponding vocal tract configuration for Maeda's synthesizer. These values are shown in Table 4 below.

Table 4: Fixed synthesis parameters based on target phoneme type

Phoneme type ID	Synthesis parameter					
	JW	TP	TS	TA	LA	LP
1	1	-1	0.6	0	0	-1
2	-0.6	-0.6	-1.8	0	0	-1
3	-0.6	0	0	0	1.1	0.6
4	-1.3	0.8	-0.5	0	2	-1
5	3	1.3	1.2	-0.3	-0.6	-0.2
6	0.3	1.3	1.2	-0.3	-0.6	-1.5
7	0.1	1.4	-0.2	-0.1	0	0
8	-3	1.5	-3	-0.1	-0.3	-2.8
11	0	0	0	0	-1.5	-3
12	0	0	0	0	-1.5	-3
13	0.4	-1.8	-0.1	-0.2	-1.1	3
14	0	0	0	0	-1.2	-3

15	0	-1	-1.8	0	-1.1	-3

While this approach is robust and can be used for any phoneme that has known target synthesis values, it suffers from several problems. The vocal synthesis is in some cases a many-to-one mapping, which means that there are multiple possible synthesizer values that could result in very similar acoustic sounds, and there is not a particular method underlying the selection of the Table 4 synthesis parameters that makes for a smoothly varying parameter space across the entire vocal space. This makes creating a linear mapping problematic.

A second issue is that these target values are prototypical and therefore do not match the acoustics of each speech instance being used to generate the kinematics, which means that the training data itself is not accurate, so that the synthesis parameters are not matched to the kinematics.

To address this issue, a second approach for generating synthesis parameters was implemented, based on acoustic matching as described in the next section.

4.3.2. Target synthesis parameters based on acoustic formant matching

The goal of this approach is to determine the corresponding synthesis parameters by matching those parameters to the acoustics of the corresponding input waveform, for each training example. For consonants it is difficult to identify a clear acoustic feature or spectral measure that can be used, so the experimental work thus far has been

limited to vowels. Vowels almost always have four or more distinguishable formants, but the first two formants are most important to determine the quality of vowels [15] . Hence, F1 and F2 play critical roles in synthesis parameter selection.

The relationship between synthesis parameters and generated formant values in the Maeda synthesis has not been thoroughly studied, and so the first step was a more in-depth study of this relationship. A sub-selection technique was then applied to identify points within the formant space that correspond to a more smoothly-varying set of synthesis parameters. From these values, formant values of the training waveforms were used to lookup appropriate target synthesis values, which were used for training the mapping.

The method used for the initial analysis of formant-synthesizer relationships was a grid-based search of combinations of the most acoustically-relevant synthesis parameters. An approximate initial range of synthesis parameters was chosen, as follows:

1. JW: from -3 to 3, the interval is 0.25;
2. TP: from -1 to 2.5, the interval is 0.25;
3. TS: from -2.5 to2.5, the interval is 0.25;
4. TA = from -0.5 to 2.5, the interval is 0.25;

Across this range of parameters, the VTCalc software, implemented in Matlab, was used to generate 100ms length vowel signals. After this, a basic formant estimation using a low-order LPC model was implemented to estimate the F1 and F2 formant values. While this is an overly simplistic formant estimation technique for real speech, for the highly

stationary synthetic speech produced by the articulatory synthesizer, the approach worked well. The LPC order used was 10, because empirically this resulted in the most consistently accurate F1 and F2 values as shown in Figure 48. The first two peaks above an initial threshold of 200 Hz were used as estimates.

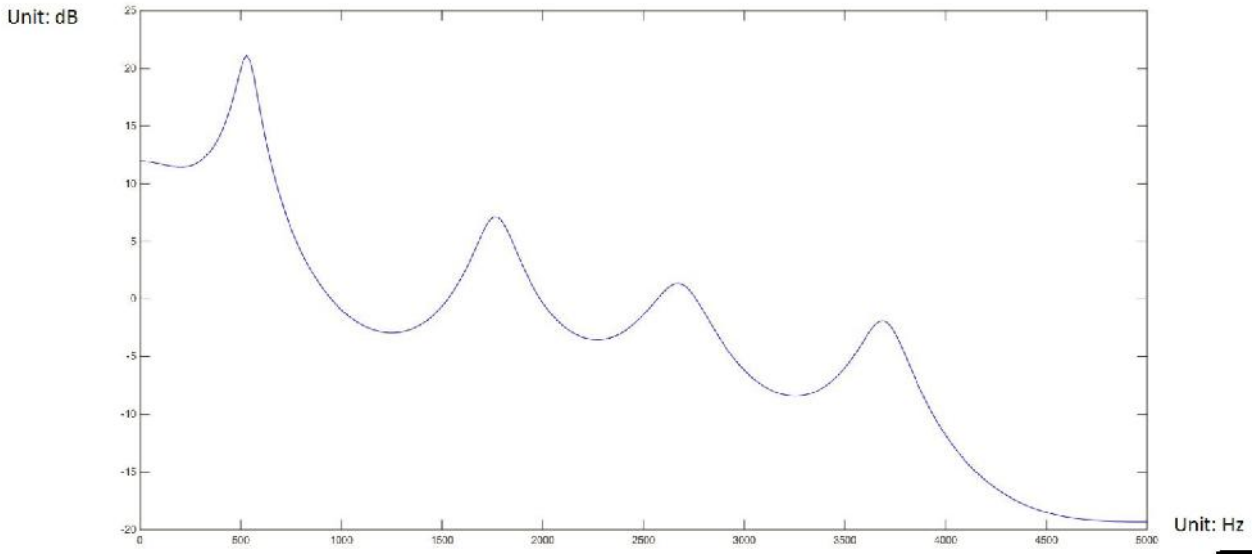

Figure 48: Example LPC spectral envelope of order 10 for a synthesized vowel /a/, used to estimate F1 and F2.

After estimating the corresponding F1 and F2 formant values across the synthesis parameter range mentioned above, we plotted the associated formant space, as shown in Figure 49

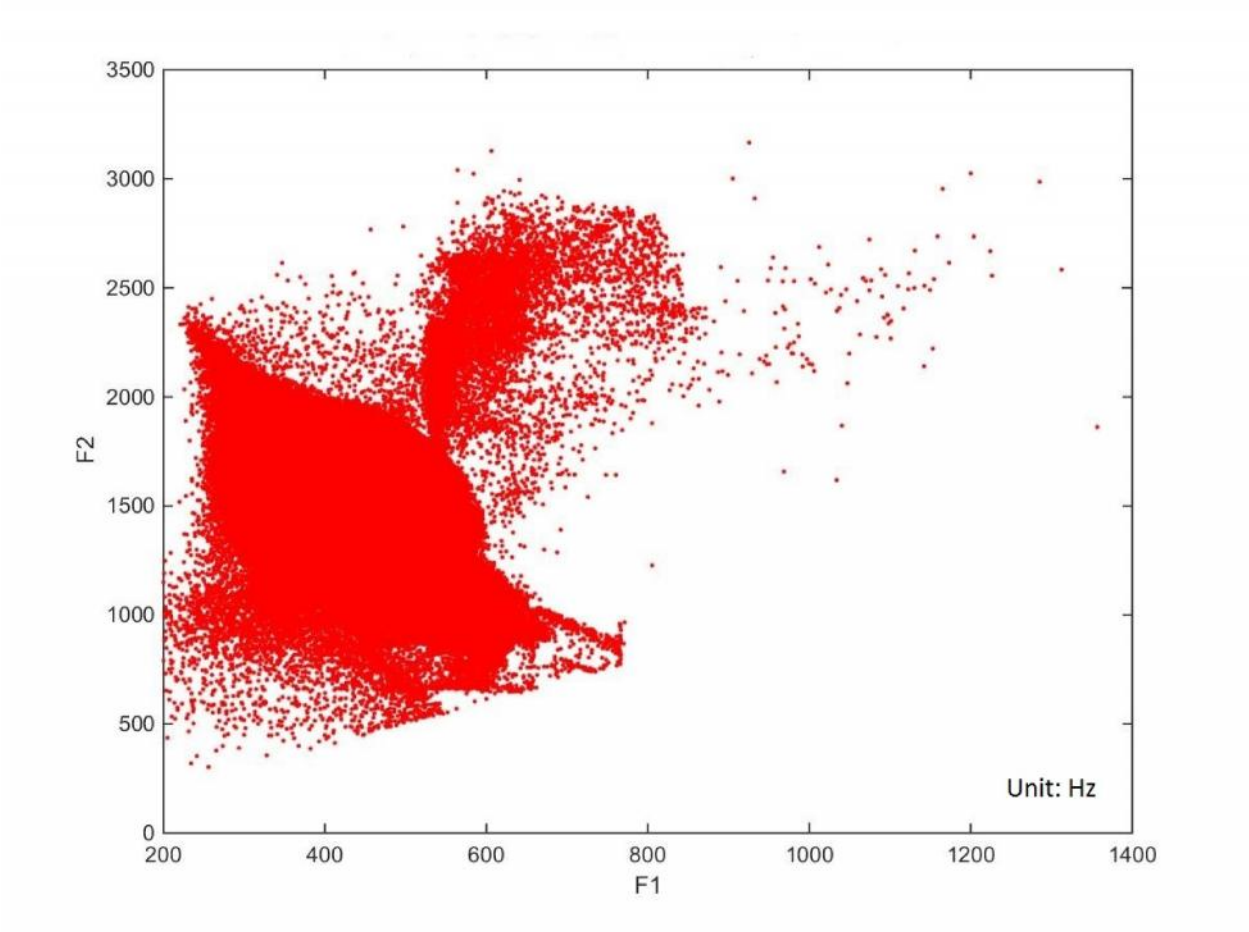

Figure 49: The distribution of vowel formants across all synthesis parameters

Because of the many-to-one characteristics of the synthesizer, there are many cases where settings that are close in terms of formant values are not close in terms of parameter settings. To address this, one approach is to sub-sample the full range of parameter settings in a way that enforces smoothness in both the formant space and the parameter space, i.e. to select points in such a way that those that are close in terms of formant values are also as close as possible in terms of synthesis parameter settings.

To accomplish this, a sub-sampling algorithm was implemented. The algorithm is described step-by-step as follows:

1. An initial set of points was selected randomly, by sampling the points in Figure 50, referred to as the formant space. The size of this subgroup was arbitrarily set at 10% of the full space. This group of points was named the “sample group”.
2. One target point from within the sample group was selected for evaluation. A set of 3-4 nearby points also in the sample group was selected based on formant distance (+/- 3Hz). This group was identified as the “neighborhood group.”
3. Around the target point, another set of 5-6 nearby points from the full formant space was selected, also based on formant distance (+/- 10Hz). This group was identified as the “swap group”. (Because this group is selected from the full group, its size in terms of formant distance is several times smaller in diameter.)
4. The target point and each point within the swap group were evaluated for average synthesis parameter distance to all points in the neighborhood group, as measured by Euclidean distance of synthesis parameters.
5. If any of the points from the swap group had a smaller average synthesis parameter distance to the neighborhood group, the minimum distance point was swapped with the target point. This new point was added to the sample group, and the old target point was removed from the sample group.
6. Steps 2 through 5 were repeated for each point within the sample group. After this the mean synthesis distance from target point to neighborhood groups was computed and evaluated.
7. This swapping evaluation process was repeated for multiple iterations until the mean synthesis distance stopped decreasing.

Figure 50 shows the sub-sampled formant distribution after implementing the above swapping algorithm.

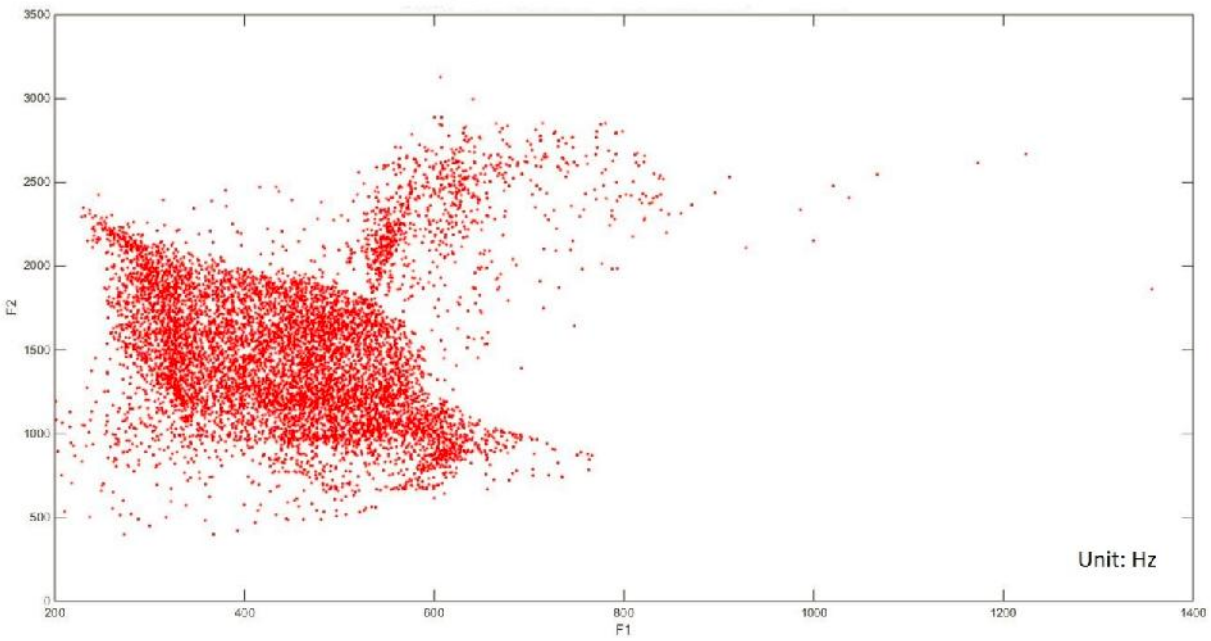

Figure 50: Target synthesis parameter distribution in the formant space

Table 5 shows the decrease in the average Euclidean distance of synthesis parameters between each point in the space and its neighbors in the formant space, as a function of the number of times the algorithm was run.

Table 5: Decrease in mean synthesis parameter distance as result of algorithm.

Previous synthesis parameter distance	Synthesis parameter distance after swapping	Swapping times
3.66	3.17	4414
3.17	2.98	2040
2.98	2.90	1049
2.90	2.86	488
2.86	2.86	163
2.86	2.85	26
2.85	2.80	0

As can be seen in the above table, the method had only minimal impact on the average synthesis parameter similarity of points in the formant space. The resulting 10% sample group was used as a table-lookup tool for matching formant values to corresponding synthesis parameters. For all the vowel training instances, the extracted segments were analyzed using a similar LPC method to identify F1 and F2. These values were also manually checked against each other and against known typical vowel formant values for consistency to identify and correct any gross estimation errors.

From the formant values for all vowel training instances, the chart illustrated in Figure 50 above was used to determine corresponding synthesis parameters. The articulatory values and final synthesis parameters were used to create the desired linear mapping, as described in the next section.

4.4. Pseudo-inverse linear method

After identifying the target synthesis parameters, we implemented pseudo inverse regression to train with articulatory features and target synthesis parameters.
A pseudoinverse A^+ of a matrix A is a generalization of the inverse matrix [16] where

$$A^+ = A^*(AA^*)^{-1}$$

A common use of the pseudoinverse is to compute a least squares solution to a system of linear equations that lacks a unique solution [17]. This provides a robust solution to the linear system $Ax = b$. Hence, through the pseudo-inverse method, we can get a

unique coefficient matrix to build the relationship between kinematic data and synthesis parameters.

This least-squares pseudo-inverse method was implemented for mapping. In the case of the synthesis parameters, there are four individual parameters being mapped, which represent four different linear mapping systems. For solving this, we did a separate least-squares pseudo-inverse for each synthesis parameter, the output of each of which is a vector of linear regression coefficients. For a specific synthesis parameter, we created a data vector of articulatory feature variables

$$X = [AF1 \cdots AF7,1]$$

The constant offset to the equation provides an additional coefficient that represents the linear regression offset.

4.4.1. The pseudo-inverse linear mapping equation

Using the matrix of articulatory features and the vector of target parameters, the linear equation for one specific synthesis parameter is given as:

$$\boldsymbol{Y} = A_{coef} \times \boldsymbol{X},$$

where Y is the vector of target synthesis parameters， A_{coef} is the coefficient vector to be solved, and X is the data matrix of articulatory features. The solution is

$$A_{coef} = X^{+} \times Y,$$

where X+ is the pseudoinverse of X.

This process is repeated for 4 times for the target variables “JW, TP, TS and TA”, respectively. (The other two target variables “UL and LL” were still computed by the 4-point linear mapping method.) After this, the four individual coefficient vectors can be combined together as a multiple-input multiple-output mapping equation as follows:

$$\mathbf{Y} = \begin{bmatrix} \mathbf{a_{11}} & \cdots & \mathbf{a_{1n}} \\ \vdots & \ddots & \vdots \\ \mathbf{a_{n1}} & \cdots & \mathbf{a_{nn}} \end{bmatrix} \times \mathbf{X}$$

This final coefficient matrix represents the linear mapping between the articulatory feature values and the synthesis parameter outputs for each synthesis parameter.

4.5. Evaluation method

To evaluate the linear mapping method, four different evaluation metrics were used. The first is an objective metric, based on the mean squared error of the synthesis parameters compared to the target valued in the training set. The second is an objective metric based on audio similarity, using formant distortion. The third is a qualitative metric, based on the Perceptual Evaluation of Speech Quality (PESQ) and the last one is based on voice distortion.

Several experimental evaluations were conducted. The first of these is the individual phoneme test, which is evaluated by MSE. The second of these is a sentence test which is evaluated by PESQ and formant distortion. In each case, the new linear mapping method is compared to the prior piecewise and quantile methods.

For mean squared error method on synthesis parameters with mapping methods, the following equation is used

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{n}\left(\hat{Y}_i - Y_i\right)^2$$

where $\hat{Y}_i$ is the value of synthesis parameter from linear mapping method or point piecewise method, and Y_i is the targeted phoneme synthesis parameter value. N is the number of samples of kinematic data.

PESQ is a standard comprising a test methodology for assessment of the speech quality. PESQ uses a perceptual model to estimate mean opinion scores (MOS) that cover a scale from 1 (bad) to 5 (excellent)[18]. It is widely used for objective voice quality and is a full-reference algorithm and analyzes the speech signal sample-by-sample. As such, PESQ makes use of an original reference signal for a comparison. In this experiment, a PESQ is used to compare the original sound file form EMA-MAE corpus, and the output signals of the Maeda synthesizer. The output signals include rebuilt sentence audio by point-wise mapping and pseudo-inverse mapping.

For formant distortion, it is used to compare the formant of original audio of sentences from EMA-MAE corpus with output signals of the Maeda synthesizer. The output signals include rebuilt sentence audio by point-wise mapping and pseudo-inverse mapping. The distortion formula is in the following:

$$\text{formant\_distortion} = \sqrt{\left((\text{F\_original}) - (\text{F\_synthesized})\right) \times \left((\text{F\_original}) - (\text{F\_synthesized})\right)^{*}}$$

4.6. Mapping experiments

4.6.1. Experimental Setup

Speech data from 5 native speakers, 3 female and 2 male in the EMA-MAE corpus, was used. The words section from five native speakers in the EMA-MAE corpus is used to measure articulatory features for training and also is used for evaluation. In the words section, there are 17 word lists and each list has 24 words, respectively. In addition, the sentence section which have 10 sentences for individuals from those 5 native speakers in the EMA-MAE corpus were implemented for evaluation.

We used the segmented phonemes from the words section of the corpus for computing the linear mapping relationship for training and also for MSE evaluation.

Experimental setup is summarized in Table 6. The individual vowels and consonants from the words section of the data as well as the continuous speech data from the corpus were used for evaluation, with metrics appropriate to those data as described in the table.

Table 6: Experimental setup. Evaluation metrics used with each data group and mapping method are indicated by an X. Data groups where the data used to determine the mapping are the same as those used for evaluation are indicated as training set data.

Data	Mapping Methods	Evaluation data class	Evaluation Method		
			MSE	Formant distortion	PESQ
Word section					
vowel	Pseudo-inverse mapping based on fixed parameters	Training set	X	X	X
	Pseudo-inverse mapping based on determined parameters	Training set	X	X	X
	4-point piecewise mapping	Training set	X	X	X
	Quantile-based mapping	Training set	X	X	X
consonant	Pseudo-inverse mapping based on fixed parameters	Training set	X		
	4-point piecewise mapping	Training set	X		
	2-Point piecewise linear mapping	Training set	X		
	Quantile-based mapping	Training set	X		
Sentence section					
	Pseudo-inverse mapping based on fixed parameters	Testing set			X
	Pseudo-inverse mapping based on determined parameters	Testing set			X
	4-point piecewise mapping	Testing set			X
	2-Point piecewise linear mapping	Testing set			X
	Quantile-based mapping	Testing set			X

4.6.2. The experimental results

Results of the experiments are shown below. Table *7* to

Table *11* show the mean-squared synthesis error on the training data for a single subject using the pseudo inverse method with target synthesis parameters chosen by phoneme type, for the synthesis parameter JW, TP, TS and TA respectively, compared to the mean-squared synthesis error using the Quantile method. Results are subdivided into vowel, consonant, and overall error. From these figures it is clear that linear mapping has better performance than quantile mapping method in most cases.

Table 7: Mean squared error for subject 36 for all synthesis parameters comparing the least-squares mapping method using synthesis parameters determined by phoneme ID to the Quantile method

Vowel				
	JW	**TP**	**TS**	**TA**
Pseudo-inverse mapping based on fixed parameters	3.66	1.27	2.51	0.27
Quantile-based mapping	5.04	1.63	4.01	0.33
Consonant				
	JW	**TP**	**TS**	**TA**
Pseudo-inverse mapping based on fixed parameters	0.15	0.71	2.93	0.28
Quantile-based mapping	0.93	2.19	2.94	0.45
Vowel and Consonant				
	JW	**TP**	**TS**	**TA**
Pseudo-inverse mapping based on fixed parameters	2.00	0.98	2.55	0.17
Quantile-based mapping	2.99	1.91	3.47	0.39

Table 8 : Mean squared error for subject 37 for all synthesis parameters comparing the least-squares mapping method using synthesis parameters determined by phoneme ID to the Quantile method

Vowel				
	JW	**TP**	**TS**	**TA**
Pseudo-inverse mapping based on fixed parameters	3.63	1.27	2.49	0.21
Quantile-based mapping	5.35	1.23	5.03	0.57
Consonant				
	JW	**TP**	**TS**	**TA**
Pseudo-inverse mapping based on fixed parameters	0.16	0.70	2.92	0.29
Quantile-based mapping	0.55	1.40	2.61	0.64
Vowel and Consonant				
	JW	**TP**	**TS**	**TA**
Pseudo-inverse mapping based on fixed parameters	2.15	1.08	2.51	0.18
Quantile-based mapping	2.95	1.31	3.82	0.61

Table 9: Mean squared error for subject 38 for all synthesis parameters comparing the least-squares mapping method using synthesis parameters determined by phoneme ID to the Quantile method

Vowel				
	JW	**TP**	**TS**	**TA**
Pseudo-inverse mapping based on fixed parameters	8.40	2.59	3.91	0.32
Quantile-based mapping	7.57	1.72	6.54	2.14
Consonant				
	JW	**TP**	**TS**	**TA**
Pseudo-inverse mapping based on fixed parameters	0.21	0.75	2.88	0.30
Quantile-based mapping	1.34	2.83	3.90	1.75
Vowel and Consonant				
	JW	**TP**	**TS**	**TA**
Pseudo-inverse mapping based on fixed parameters	2.02	0.98	2.43	0.16
Quantile-based mapping	2.84	3.54	2.95	1.82

Table 10: Mean squared error for subject 39 for all synthesis parameters comparing the least-squares mapping method using synthesis parameters determined by phoneme ID to the Quantile method

Vowel				
	JW	**TP**	**TS**	**TA**
Pseudo-inverse mapping based on fixed parameters	1.01	2.84	1.08	0.19
Quantile-based mapping	4.34	4.25	1.99	1.90
Consonant				
	JW	**TP**	**TS**	**TA**
Pseudo-inverse mapping based on fixed parameters	0.70	6.83	1.79	1.30
Quantile-based mapping	0.51	0.95	3.18	0.40
Vowel and Consonant				
	JW	**TP**	**TS**	**TA**
Pseudo-inverse mapping based on fixed parameters	2.01	1.03	2.49	0.20
Quantile-based mapping	4.04	1.34	4.86	1.27

Table 11: Mean squared error for subject 40 for all synthesis parameters comparing the least-squares mapping method using synthesis parameters determined by phoneme ID to the Quantile method

Vowel				
	JW	**TP**	**TS**	**TA**
Pseudo-inverse mapping based on fixed parameters	3.53	1.44	2.71	0.18
Quantile-based mapping	4.23	1.36	4.35	0.37
Consonant				
	JW	**TP**	**TS**	**TA**
Pseudo-inverse mapping based on fixed parameters	0.32	0.73	4.75	0.65
Quantile-based mapping	0.31	1.24	3.22	0.50
Vowel and Consonant				
	JW	**TP**	**TS**	**TA**
Pseudo-inverse mapping based on fixed parameters	1.93	1.04	2.51	0.19
Quantile-based mapping	2.27	1.30	3.79	0.43

Table 12 shows the corresponding average value of the mean squared error of synthesis parameters comparing the least-squares method using phoneme ID target synthesis parameters to other comparative methods for vowel data, for 5 subjects. Note that the reference value used for calculating mean-squared error is the same target synthesis value used for the least-squares approach, which biases the results toward the proposed method.

Table 12: The average mean squared error for various mapping methods on vowel data, using target synthesis parameters based on phoneme ID.

	JW	**TP**	**TS**	**TA**
Pseudo-inverse mapping based on fixed parameters	5.85	1.87	4.47	0.72
4-point piecewise mapping	5.09	3.31	5.57	1.53
2-Point piecewise linear mapping	5.2	3.86	5.75	1.56
Quantile-based mapping	5.30	2.04	4.38	1.66

Table 13 shows the corresponding average value of the mean squared error of synthesis parameters determined by formant matching with vowel data, for 5 subjects.

Table 13: The average mean squared error for various mapping methods on vowel data, using target synthesis parameters based on parameters determined by formant matching

	JW	**TP**	**TS**	**TA**
Pseudo-inverse mapping based on parameters determined by formant matching	2.50	1.08	2.05	0.66
4-point piecewise mapping	2.52	1.25	2.16	0.90
2-Point piecewise linear mapping	2.71	1.74	2.70	1.29
Quantile-based mapping	3.96	1.98	2.74	3.42

Table 14 shows the average PESQ and formant distortion for the vowel phoneme which used the audio signal generated by synthesis parameters from mapping methods.

Table 14: PESQ and formant distortion for vowel data

	PESQ	Formant distortion
Pseudo-inverse linear mapping based on fixed synthesis parameters	2.88	360.51
Pseudo-inverse mapping based on parameters determined by formant matching	2.93	336.28
4-point piecewise mapping	2.46	398.13
2-Point piecewise linear mapping	2.14	463.25
Quantile-based mapping	2.33	417.02

Table 15 shows the corresponding average value of the mean squared error of synthesis parameters based on fixed synthesis parameters with consonant data.

Table 15: the mean squared error for mapping method based on fixed synthesis parameters with consonant data

	JW	**TP**	**TS**	**TA**
Pseudo-inverse mapping based on fixed parameters	2.20	0.92	1.97	0.59
4-point piecewise mapping	2.52	1.25	2.16	0.90
2-Point piecewise linear mapping	2.71	1.74	2.70	1.29
Quantile-based mapping	0.73	1.72	3.17	0.75

Table 16 shows the average value of PESQ on the sentence section.

Table 16: the results of PESQ on sentence

	PESQ
4-Point piecewise linear mapping	1.92
2-Point piecewise linear mapping	1.83
Quantile-based linear mapping	1.74
Pseudo-inverse linear mapping based on fixed synthesis parameters	1.22
Pseudo-inverse linear mapping based on formant matching parameters	1.32

4.7. Discussion and Conclusions

This chapter has introduced a new linear mapping method based on the pseudo-inverse method, using two different approaches for establishing synthesis targets. The method has been implemented on five native speakers in the EMA-MAE corpus.

Results of the new method were mixed, but were successful in revealing several useful insights about both the current method and potential of the new approach. Overall conclusions include:

- Using target synthesis parameters based on phoneme ID resulted in significantly higher mean-square error than using target parameters based on acoustic matching. This suggests that using fixed phoneme-based targets may not be the best approach.
- The relatively high error even when using formant-based targets suggests the possibility of subject-synthesizer acoustic mismatch. Because the VTdemo software is based on specific atypical and non-native subjects, effort spent matching the vocal space of the subject and synthesizer may have substantial value.
- PESQ assessment was not very useful. Because the PESQ requires a reference "clean" signal which this application does not provide, it only gave very rough measures of quality. Better ways to measure quality of continuous speech are needed.

- Overall error levels suggest the need for more exemplars in building the least-squares mapping, which requires clear methods to map synthesis parameters to acoustics.
- The pseudoinverse mapping approach using formant-driven synthesis targets demonstrated lower error for nearly every synthesis parameter and phoneme type. This suggests that a subject-matched least-squares mapping is likely to be an improvement over previous mapping methods.

These overall conclusions suggest several specific areas for future work:

- New methods to match the vocal space and acoustic characteristics to those of the synthesizer, including either additional mapping (e.g. pre-mapping formant adjustments) or synthesizer modifications (e.g. synthesizer scaling factors). If this can be done automatically on continuous speech, significant improvements may be expected.
- New methods to assess vocal quality of continuous speech

CHAPTER 5 Conclusion

This thesis has introduced a least-squares linear mapping method for accurately mapping articulatory kinematic data from an EMA system onto acoustic synthesis parameters. In order to realize this new linear mapping approach, kinematic features based on a three dimensional palate mesh were used to provide an initial input more representative of the vocal tract structure, and a new approach for determining accurate synthesis parameter targets based on formant value matching was introduced. Experimental results on several subjects from the EMAMAE dataset indicate that the new mapping gives reduced mapping error. Ultimately, the impact of this work is that it provides researchers with a more accurate method for mapping kinematic data to synthesis parameters.

5.1. Summary of work

A gridded convex hull method for analysis and estimation of palate trace has been developed and implemented for the creation of a virtual 3D palate trace. In addition, a methodology for determining acoustically matched synthesis parameters for training a mapping has been developed. The primary work consists of the implementation of the pseudoinverse method for linear mapping, and the implementation of several different prior mapping methods for comparison. Overall, the acoustic feedback from novel

mapping method is improved and the distortion and delay from previous methods are decreased.

5.2. Research Contributions

This thesis includes three major contributions:

1. The creation of a new approach for estimating the three dimensional virtual palate trace for individual subjects, based on a gridded convex hull and thin-plate spline.
2. The creation of a new approach for estimating correct target synthesis values for kinematic training data, using a formant-based acoustic matching algorithm.
3. The application of the pseudoinverse linear mapping method to the problem of kinematic-to-synthesis parameter mapping.

Based on the experimental results, the kinematic to synthesis mapping method is able to estimate more accurate synthesis parameters. This will enable the RASS system to provide acoustic feedback to subjects with less distortion.

5.3. Future Work

There are several directions for improving the kinematic to synthesis linear mapping. In addition to those ideas discussed in Chapter 5, another valuable extension of this work would be to create the virtual three dimensional palate trace based on more accurate palate information, for example by incorporating MRI scans or other imaging

technology. It would also be possible to extend the number of articulatory features which can illuminate vocal tract movement during speech. Moreover, more regression methods including nonlinear mapping techniques could be incorporated into the mapping process.

Bibliography

[1] B. E. Murdoch and G. T. Deborah, "Traumatic brain injury: Associated speech, language, and swallowing disorders," in San Diego:Singular/Thomson Learning, 2001, .

[2] Huckvale, "VTDemo-Vocal Tract Acoustics Demonstrator," 2009.

[3] S. Maeda, "Compensatory Articulation during Speech: Evidence from the Analysis and Synthesis of Vocal-Tract Shapes using an Articulatory Model," *Speech Production and Modelling,* pp. 131-149, 1989.

[4] J. S. Perkell, "Movement goals and feedback and feedforward control mechanisms in speech production," *Journal of Neurolinguistics,* vol. 25, pp. 382-407, 2012.

[5] T. Asterios, O. Slim and L. Yves, "Estimating the Control Parameters of an Articulatory Model from Electromagnetic Articulograph Data," *Journal of the Acoustical Society of America,* vol. 129, pp. 3245-3257, 2011.

[6] J. Berry, "Accuracy of the NDI wave speech research system," *Journal of Speech-LanguageHearing Research,* pp. 1295-1301, 2011.

[7] J. S. Perkell, H. Lane, M. Svirsky and J. Webster, "Speech of cochlear implant patients: a longitudinal study of vowel production," *J. Acoust. Soc. Am. 91,* pp. 2961–2978, 1992.

[8] A. Ji, M. T. Johnson and J. Berry, "Tracking articulator movements using orientation measurements," in *International Conference on Audio, Language, and Image Processing,* Shanghai, China, 2012, .

[9] Yana Yunusova, Melanie Baljko, Grigore Pintilie, Krista Rudy, Petros Faloutsos and John Daskalogiannakis, "Acquisition of the 3D surface of the palate by in-vivo digitization with Wave," *Speech Communication,* pp. 923-931, 2012.

[10] C. Bradford Barber, D.P. Dobkin and H. Huhdanpaa, "The quickhull algorithm for convex hulls ACM Trans," *Math. Software, 22 (4),* pp. 469–483, 1996.

[11] Branko Grünbaum, *Convex Polytopes.* 2003.

[12] T. Cormen H, C. Leiserson E and R. L. Rivest, "Introduction to Algorithms," (MIT Press, 2000).

[13] J. Laver, Ed., *Principles of Phonetics.* Cambridge University Press, 1994.

[14] S. Maeda, "A digital simulation of the vocal-tract system," *Speech Commun.,* vol. 1, pp. 199-229, 1982.

[15] D. Deterding, "The Formants of Monophthong Vowels in Standard Southern British English Pronunciation," *Journal of the International Phonetic Association,* pp. 47-55, 1997.

[16] Thomas N.E. Greville, *Generalized Inverses.* Springer-Verlag, 2003.

[17] R. Penrose, "On best approximate solution of linear matrix equations," *Proceedings of the Cambridge Philosophical Society 52,* pp. 17–19, 1956.

[18] ITU, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," *ITU-T Rec. P. 862,* 2000.