Using Synchronized Audio Mapping to Track and Predict Velar and Pharyngeal Wall Locations during Dynamic MRI Sequences

Pooya Rahimian[1], Jamie L. Perry[2,*], Lakshmi Kollara[3] and Nasseh M. Tabrizi[4]

[1]*Department of Computer Science, University of Iowa, Iowa City, IA, 52240, USA*

[2]*Department of Communication Sciences and Disorders, 3310 Allied Health Sciences, East Carolina University, Greenville, NC 27834, USA*

[3]*Department of Communication Sciences and Disorders, 3310 Allied Health Sciences, East Carolina University, Greenville, NC 27834, USA*

[4]*Department of Computer Science, East Carolina University, Greenville, NC 27834, USA*

Abstract: *Purpose*: The purpose of this study is to demonstrate a novel innovative computational modeling technique to 1) track velar and pharyngeal wall movement from dynamic MRI data and to 2) examine the utility of using recorded participant audio signals to estimate velar and pharyngeal wall movement during a speech task. A series of dynamic MRI data and audio acoustic features were used to develop and inform a Hidden Markov Model (HMM) and Mel-Frequency Cepstral Coefficients (MFCC) model.

Methods: One adult male subject was imaged using a fast-gradient echo Fast Low Angle Shot (FLASH) multi-shot spiral technique to acquire 15.8 frames per second (fps) of the midsagittal image plane during the production of "ansa." The nasal surface of the velum and the posterior pharyngeal wall was identified and marked using a novel pixel selection method. The error rate was measured by calculating the accumulation error and through visual inspection.

Results: The proposed model traced and animated dynamic articulators during the speech process in real-time with an overall accuracy of 81% considering one pixel threshold. The predicted markers (pixels) segmented the structures of interest in the velopharyngeal area and were able to successfully predict the velar and pharyngeal configurations when provided with the audio signal.

Conclusion: This study demonstrates a novel and innovative approach to tracking dynamic velopharyngeal movements. Discussion of the potential application of a predictive model that relies on audio signals to detect the presence of a velopharyngeal gap is discussed.

Keywords: Hidden Markov Model, dynamic MRI, velopharyngeal position, computational modeling, Mel-Frequency Cepstral Coefficients model.

INTRODUCTION

The most commonly used direct assessments for visualizing the velopharynx are nasoendoscopy and multi-view videofluoroscopy [1-4]. Nasoendoscopy is invasive and often provides distorted depth cues limiting interpretations related to size of the structures and the extent of closure [5-7]. The wide angle distortion and oblique angle of view can produce errors in the validity and reliability of estimates of the nasopharyngeal orifice, particularly related to lateral pharyngeal wall movement [8, 9]. Videofluoroscopy uses ionizing radiation, thus limiting the safety of prolonged and repeated use for assessment and follow-up and can present with errors particularly due to misalignment of the patient and measurement error [10]. The most notable limitation of current direct visualization techniques is the inability to quantify velopharyngeal activity [5, 11, 12]. As such, research continues to emphasize the need for a clinical dynamic tool for assessing velopharyngeal function.

Dynamic magnetic resonance imaging (MRI) can provide valuable information regarding velopharyngeal structures during speech. The advantages of dynamic MRI include the ability to specify the exact plane of interest for imaging which eliminates the depth perception distortions found in nasoendoscopy and the ability to accurately calculate the in-plane orifice size and velopharyngeal gap [13, 14]. Studies have demonstrated the use of non-cyclic dynamic MRI in children and adults at frames rates between 15.8 and 100 frames per second [14-16]. These MRI protocol are designed to be independent of repetitions, be acquired rapidly, and allow sentence-level productions to promote more natural speech acquisition [17]. Traditionally, manual tracings have been used for image segmentation of the velopharyngeal structures [17, 18]. These methods are extremely time consuming and may demonstrate inter-rater variability. Studies have been conducted using ultrasound, x-ray, and MRI

*Address correspondence to this author at the Department of Communication Sciences and Disorders, 3310Q Allied Health Sciences, East Carolina University, Greenville, NC 27834, USA; Tel: (252) 744-6144;
Fax: (252) 744-6104; E-mail: perryja@ecu.edu

E-ISSN: 2311-1917/16

to examine the dynamic nature of the velopharyngeal structures during speech. Noise, motion artifacts, air interfaces, and refractions often complicate the process of computer-based automatic tracings. One method to overcome the errors associated with computer-based tracing algorithms is to use patient data to first train the model on the kinematic details. Many image segmentation algorithms work with some prior knowledge regarding the shape and/or location of target objects.

Speech recognition and segmentation algorithms are part of a multi-disciplinary area of research in which many studies have been performed [19]. These studies primarily use Mel-Frequency Cepstral Coefficients [20, 21] to extract audio features to improve quality for extracting features of the human voice [20, 22]. Mel-Frequency Cepstral Coefficients has been widely used in music analysis studies. To our knowledge, no studies have provided a method for tracking dynamic MRI data of the velopharyngeal mechanism during speech production. This study demonstrates the use of MFCC to extract audio features and combined audio and visual features were feed a supervised Hidden Markov Model (HMM) to track velopharyngeal movements from dynamic MRI data.

Within this study, we also demonstrate a novel method for reversing the segmentation algorithm to examine if acoustic information (recorded speech) could be used to predict the velar and pharyngeal locations, thus identify a presence or absence of a velopharyngeal gap. Following the development of the HMM, we trained the model to evaluate if it could predict the location of velar and pharyngeal structures based on the audio signals. These innovations may be particularly useful for school-based speech pathologists by providing a cost-efficient instrument to confirm their perceptual speech assessments and provide support for appropriate referrals to cleft palate craniofacial teams.

METHOD

Subject

In accordance with the local Institutional Review Board, one healthy adult (21 years old) male subject was recruited to participate in the study. The subject had normal speech, language and hearing and indicated no history of a craniofacial anomaly, swallowing disorder, sleep apnea, or neurologic disorder. The subject was scanned in the supine position using a Siemens 3 Tesla Trio (Erlangen, Germany) MRI scanner and a 12-channel Siemens Trio head coil. Simultaneous speech recordings were obtained following previously described methods [3, 4, 6]. The subject wore an MR-compatible headset with an attached optical microphone (Dual Channel-FOMRI, Optoacoustics Ltd., Or Yehuda, Israel). The optical microphone has two channels for active cancellation of the loud MR gradient noise while preserving the speech sample from the subject.

The dynamic MRI protocol has been previously described [2, 3] and includes a fast-gradient echo Fast Low Angle Shot (FLASH) multi-shot spiral technique was used to acquire 15.8 frames per second (fps) of the midsagittal image plane during the production of “ansa.” The speech sample was chosen to represent movements of the velum between fully lowered (i.e., nasal), elevated (i.e., consonants), and transitions between both positions. A metronome beat of 2 Hz was played over the head phones to control the rate of the speech tasks (two syllables per beat). This imaging speed allowed for at least one full image during each lowered and each elevated production to analyze the data for a nasal and oral sound.

Images were reconstructed with an output time-driven sliding window process at 40 frames per second (fps). This process allowed data to have a minimal amount of interpolation across time and uses the native frame rate (15.8 fps) to interpolate images to the desired output rate. The sliding window reconstruction process minimized redundant information in adjacent time points and minimizes temporal blurring [23]. Acquisition simulation software provided by the vendor of the MRI scanner provided timing data which was used to align the audio speech recording with the dynamic images. This software allowed for accurate simulations of sequence timing using the exact acquisition protocol, providing information about data acquisition events with 10 μs accuracy.

Dynamic MRI movies were imported into a visual and motion graphic software program (Adobe After Effects, CS 6, Adobe Systems) where data were exported as image sequence of the entire 45 seconds. A 7.5 second segment of the 45 seconds was selected. The audio and image sequences were isolated from 7.5 seconds of video at 16 KHz and 40 frames per second respectively in order to produce a number of frames per second to be a factor of 1:400 sampling rate. A ratio of 1:400 creates the proper window size for audio feature extraction phase because MFCC requires

consecutive samples per frame in order to reflect corresponding audio features correctly. If the size of window is too short, MFCC cannot extract features while larger window sizes may have adverse impacts on training time and makes an unnecessary complex model. This ratio was selected experimentally.

The MRI device records images on a constant frame rate, thus capturing the trajectory of the velum in consecutive discrete frames. A similar discretized approach is applied in the audio signal recording. The primary characteristic of these two data are that they are temporal events, in which chronology of occurrence is a main factor (i.e., the data has an order based on time). HMM is similar to the aforementioned data with respect to timing aspects. The order of occurrence of an event (i.e., timing aspect) is a factor in the prediction process in HMM. In the training phase for the model, audio and corresponding visual features are fed into the HMM in chronological order. HMM estimates the possibility of marker appearance for each possible location based on the given audio features, visual features, and previous audio features. Hypothetically, there are many locations (i.e., any pixel) that can represent a marker at a given time. However, the likely locations of markers are in fact a variable of their previous locations. Thus the possibility of those pixels being close to the previous location is higher than them being in other locations. After the training phase is completed, the HMM is able to predict the location of markers by analyzing audio features in the absence of visual clues, because it was tuned by a set of dummy data. The model can thus locate the most likely position for each marker at any given time, post training.

Audio Feature Extraction

Noise removal was accomplished by passing the original audio signal through a multi-band noise gate by using a 0.5 seconds segmented noise sample profile. Spectral noise gating [24] algorithm considers the given sample noise as the noise floor (or threshold). Fast Fourier Transform (FTT) was then applied to the audio sample for each band of the spectrum. The noise is then classified for each frequency band by finding the maximum value. The original signal is compared with corresponding threshold gate to determine whether it should be passed or discarded. Signals greater than the gate threshold pass as well as the remaining signals are discarded by the corresponding gate. The noise gate filter is sensitive to the sample profile. Therefore, 0.5 seconds of silence was segmented from the immediately prior to the actual speech signal (Figure **1**). After noise reduction, the audio signal feature extraction was accomplished by using MFCC. The MFCC are short-term spectral-based features [20] where the mel-scale is chosen close to the human auditory system [25].

Feature discretization, as seen in Figure **1** (box three, top image), is a process of converting the continuous audio features into discrete numbered groups (i.e., features) which are then applied to the HMM for the purpose of training. In this study, the discrete groups represented an audio feature that was paired with velar and pharyngeal wall shapes. The final acoustic feature dimension was 39 elements including MFCC coefficient transformations (with 13 elements) and the first and the second derivatives (13 elements for each derivative). The extracted features are similar to the audio signal, and consist of a discrete stream of features. Length of feature stream was divided by number of video frames to find corresponding audio features per frame. In this study, 39×400 features represented one frame of video. HMM accepts discrete features; therefore, features were discretized and labeled in 400 distinct classes from 1 to 400 (i.e., rounding up features). In such, we were able to train the computerized model to enable accurate prediction the location and shape of the velum and pharyngeal

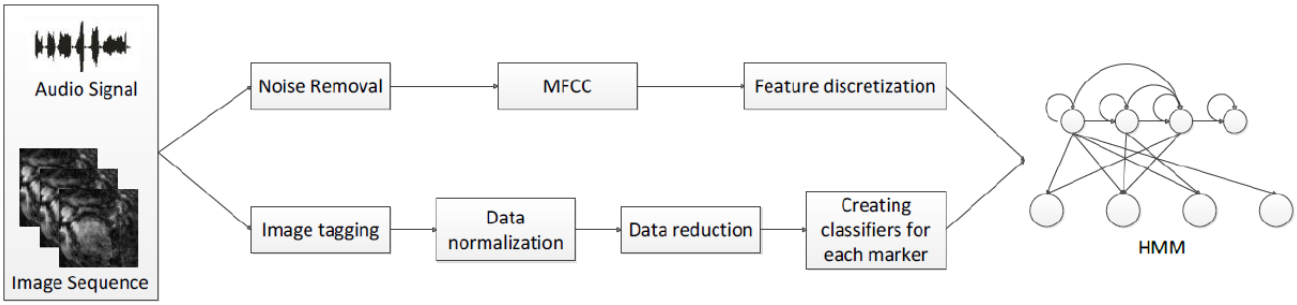

Figure 1: In order to train the model, HMM needed two data sets consisting of audio and visual data sets. Raw audio signal was passed through top pipeline to extract audio features and corresponding image sequence was tagged through the bottom pipeline. Both audio and visual features were fed into the HMM to accomplish training phase.

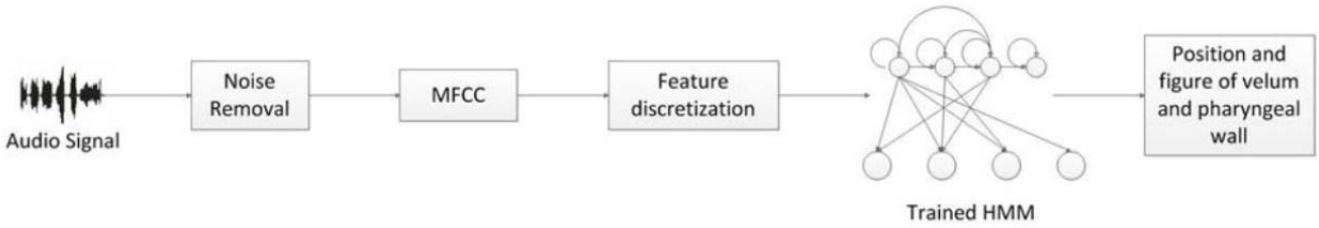

Figure 2: Trained HMM is able to predict the location of the velum and pharyngeal wall solely based on the given audio signal. Predicted position is a set of 2D vectors illustrating location and figure of velum and pharyngeal wall superimposed on magnetic resonance images.

wall based on given audio signal. (i.e., production of /ansa/). As shown in Figure **2**, the pipeline of audio feature extraction was cloned for the prediction phase. Three consecutive blocks of audio feature extractions convert the audio signal into a stream of features and the trained HMM is then able to predict the location of structures for every window (39×400). This process is completely automated and the result of prediction can be superimposed on MRI data.

Visual Feature Extraction

Visual features were extracted using the MR image sequence by selecting four markers along the nasal surface of the velum and three markers along the posterior pharyngeal wall. One stationary pivot point was placed at the posterior nasal spine (PNS). As shown in Figure **3**, the markers were positioned such that the markers were located along the length of the nasal surface of the effective velar length and not continued to the uvula proper due to lack of significance of this region during speech production.

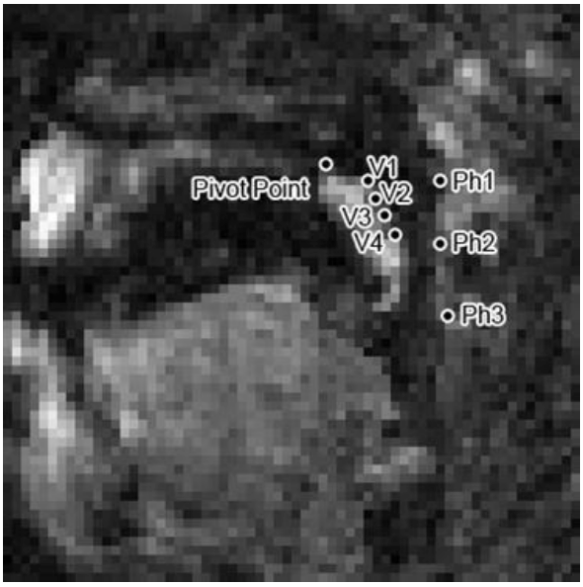

Figure 3: Demonstration of a midsagittal MR image plane showing the velar (V1-V4) and posterior pharyngeal wall (Ph1-Ph3) markers.

Equal distance between each marker was preserved using a novel circular tracking tool to identify markers along the length of the velum that extended beyond the stationary pivot point at the PNS. A circle with a 26 pixels radius was first drawn around the initial stationary pivot (i.e., PNS pivot marker). Thus, any pixel on the circumference of the circle is at an equal distance from the center of the circle. The first marker was positioned at the point of intersection of the circle with the nasal surface of the velum. The circular tracking tool subsequently creates another circle around the selected marker to identify the next marker on surface of the velum. Each marker represented one pixel and the radius around each marker was set at 13 pixels in order to achieve more markers along the length of the velum. This method, as shown in Figure **4**, provided a consistent approach in the identification of every positioned marker along the velar surface. The proposed method can be adopted for consistent placement of markers in any image of the velum.

Anterior and posterior pharyngeal wall movements were calculated in the horizontal (x-axis) dimension (Figure **5**). A reference line was drawn through the hard palate passing through the posterior pharyngeal wall (Figure **5**). The first pharyngeal wall marker (Figure **5**) was placed at the level of the hard palate. The second and third lines were placed 42 and 90 pixels below the first line, respectively. These markers were determined as the most relevant portion of the nasopharynx involved in velopharyngeal closure.

Three hundred sequential images (7.5 seconds of selected audio/video at 40 frames per second) were manually tagged by the researcher resulting in a table that consisted of 300 tuple (rows) which included 7 markers (4 velar and 3 pharyngeal) and 14 columns. Manual tagging was used to examine the validity of the HMM predictions compared to manual tagging of velopharyngeal structures. Each marker demonstrates movement in both the x and y-axis, yielding two values for each marker. For each marker, the x-value was

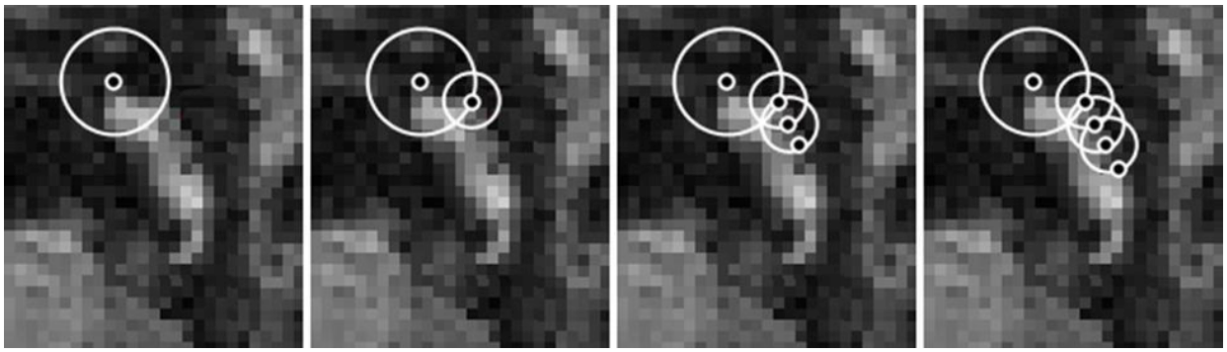

Figure 4: Demonstration of the tagging system used for determining the placement of each marker. In the far left image, the white dot is placed on the PNS. A 26 pixels radius is drawn around the pivot point (PNS) and a 13 pixels radius is drawn around each consecutive marker. The next marker is then placed at the point where the circle crosses the nasal surface of the velum.

multiplied by 1000, in order to shift the x-value to the left side and the corresponding y-value was added to create a concatenation of the x and y columns. This new set of numbers was labeled from 1 to n, which represented the total distinct classes for the compounded location of the x and y-axis for the marker across the speech sample image sequence.

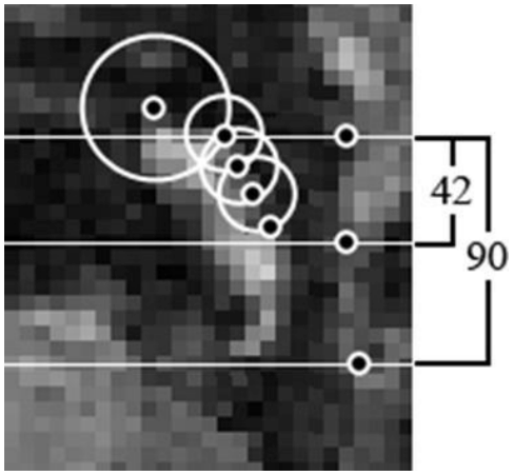

Figure 5: The superior-most pharyngeal marker was positioned along the hard palate plane. Pharyngeal markers 2 and 3 were positioned 42 and 90 pixels below the initial pharyngeal marker.

The combination of the circular tagging method and data reduction contributed to a drastic decrement in the number of hidden states. As shown in Table **1**, the model with the largest number of hidden states consisted of 18 different positions. This means that HMM eventually will designates the most probable state (i.e., position) base on given audio signal and its prediction cannot be beyond these 18 possible positions. Data reduction occasionally results in the loss of certain useful information. However, in this study, data reductions were achieved by merging less frequent inserted instances resulting from human error or low-quality of the images. Moreover, the data reduction improved the performance of the prediction system in terms of response time, because the generated model had less hidden states.

Computer Prediction Algorithm

After the audio and visual features were extracted and the HMM [26] was trained by corresponding audio and visual features, the model was used to predict velar and pharyngeal wall boundaries relative to the given audio signal. The models were trained using a 200 audio feature data set and 100 samples were set to test and evaluate the accuracy of proposed model. The internal structure of HMM is divided into two set of graphs which can be presented as two different matrices consisting of observations of the model and emission. Observation matrix is a window of extracted audio signal features (400 samples labeled 1 to 400). In the observation matrix, the possibility of transition from one label at time (*t)* to another possible label at *t*+1 (precisely from one sample to another possible sample) is stored in this matrix. Consequently, this matrix can address any arbitrary audio signal by a 400×400 matrix. Each entry in this matrix is less than or equal to one. Emission matrix is the intersection of each element in observation matrix and all possible visual features. Each entry of emission matrices, from *t* to *t*+1, represents distribution of observation values

Table 1: Number of Hidden States for Markers where V = Velar Markers and Ph = Pharyngeal Wall Markers

	V1	V2	V3	V4	Ph1	Ph2	Ph3
Hidden States	4	8	18	15	3	2	2

from time to time. In contrast to traditional linear left to right HMM often used in speech recognition systems [22], the topology of the model developed in this study did not follow a linear pattern. The model used in this study had two parameters that consisted of transition and emission matrices [26] that were estimated based on the visual and audio features. Transition and emission matrices are beneficial in that they allow for the probability of moving structures from one state to another in a dynamic system, as opposed to a system with categorical organization. This is particularly important in the velopharyngeal system, which is a very rapid and dynamic speech system. Two hundred audio samples were used to train the HMM designed in this study. One hundred audio samples (2.5 seconds or /ansa/ produced 2.5 times) were used to test the ability of the prediction model. The Viterbi algorithm [27] was then applied to predict the most likely sequence of hidden states.

RESULTS

This study demonstrates a method for using HMM and MFCC to track velar and pharyngeal wall movement during dynamic MRI data. A second aim was to examine if the novel computerized protocol could accurately identify the location of the velar and pharyngeal wall locations based on the audio signal. The accuracy of the computerized model was analyzed using two distinct methods, including *accumulative Euclidean distance* and *visual inspection*. Accumulative Euclidean distance is a mathematical approach to accumulate minimum distances between prediction and corresponding actual markers manually tagged by the researcher. The level of accuracy was evaluated as a one pixel threshold, in which an error in the model was indicated by the prediction phase producing a marker that was greater than one pixel of the corresponding marker identified by the researcher.

Accumulative Euclidean distance

This measurement introduced the accumulation of minimum distance between predicted point by trained HMM and actual markers that the researcher had placed on the image.

$$Accumulative\ Euclideance = \sum_{m=1}^{markers} \left\| L_p^m - L_r^m \right\|^2$$

Where location of predicted marker was $L_p = (x_p, y_p)$ and the location of marker was tagged by research was $L_r = (x_r, y_r)$. Zero error would indicate that every predicted marker was exactly overlaid on its corresponding manual tag. One pixel was set as the threshold because those markers having less than one pixel of error may not be visible in the superimposed image.

Figure **6** demonstrates an accumulative graph of error values for the four velar markers, where the vertical axis (top graph) demonstrates the sum of error per pixels through the velar markers and the x - axis represents the time (milliseconds). The middle and lower graphs represent a spectrogram and spectrograph respectively, for 2.5 consecutive subject productions of /ansa/. As seen in Figure **6**, the greater the amplitude, such as during the production of /a/, the greater the amount of error noted in the prediction model. The productions of /n/ and /s/ showed similar error rates which were influenced by the adjacent /a/ production (coarticulatory effect). Figure **6** demonstrates the error of the velar prediction using the computerized predictive model. Each point in the graphs accounts for the average error per pixel of each marker. Given 1 pixel threshold for these data, no error was introduced in V1 prediction, while V2, V3, and V4 were predicted with a different level of error. Error rates between the predicted and actual locations were within 1-2 pixels across velar markers. The error in V1 was dramatically less than other markers located on the velar surface. There were four hidden states (i.e., possible positions) defined for V1 (hidden states in Table **1**). Markers V2, V3, and V4 were defined by 8, 18, and 15 hidden states, respectively. Fewer hidden states resulted in less error in the prediction model. Markers V3 and V4 showed the largest number of hidden states (i.e., possible velar positions), which is a result of this region (near the velar knee) being the most dynamic, moveable segment of the velopharyngeal system. Overall, the model successfully traced and animated dynamic articulators during speech production with an overall accuracy of 81% considering one pixel threshold.

Due to limited movement of the pharyngeal wall, the residual prediction introduced a maximum of one pixel error; consequently, there was no error displayed in pharyngeal wall superimposed version. Based on the MRI resolution of .63 mm, one pixel is equal to .63 mm. At rest, the velopharyngeal port measured from the velar eminence to the posterior pharyngeal wall measured 6.6 mm (i.e., 10.4 pixels). The error rate of 1-2 pixels across markers demonstrates a highly predictive model for velar and posterior pharyngeal wall positioning.

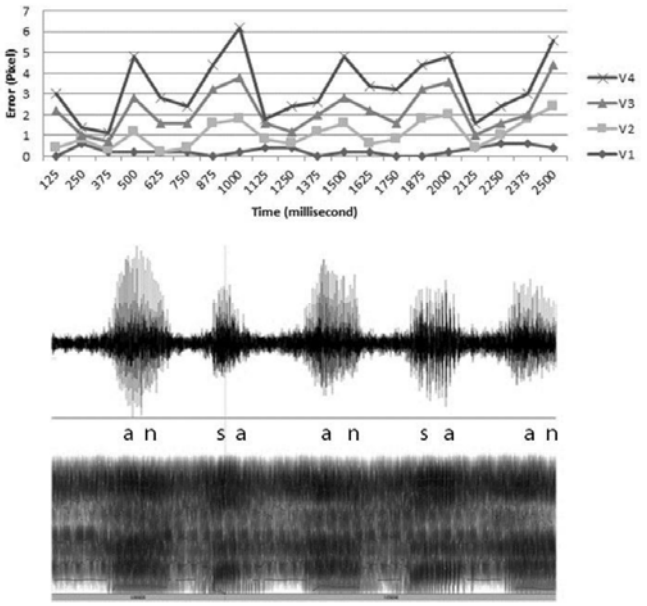

Figure 6: Velum prediction accumulative error and spectrogram with formants.

Visual Inspection

Visual inspection was performed by superimposing the predicted positions from the model over the manual markers placed by the researcher (Figure **7**). Results across each marker for every image sequence was evaluated and rated at either *pass* or *fail* assuming the same 1 pixel error rate. The result of inspection was 83% acceptance.

The inspection percentage accounts for the number of predicted frames on superimposed MR images that were accepted by the researcher. Visual inspection percentage was slightly higher (improved) than the accumulative Euclidean distance accuracy because although some of predictions had not been exactly overlaid on tags, they still laid on the boundary of the velum. Therefore, the researcher accepted these velar boundary tags as correct predictions. The difference between visual inspection and accumulative distance is not significant ($p > 0.05$) for this image set. However, we might expect divergence between these two measurements on higher resolution images, because there are more pixels potentially acceptable but not tagged.

DISCUSSION

This study demonstrates a novel and innovative methods to use computer training algorithms to track the dynamic velar and pharyngeal wall movements from MRI data. The model showed an overall accuracy of 81% considering one pixel threshold in being able to track the velopharyngeal movements from dynamic MRI data. The average error across V1 marker was 0.25 pixels, V2 was 0.835 pixels, V3 was 1.12 pixels, and V4 was 1.1 pixels. Thus, the regions of the velum that demonstrated the greatest and fastest movement trajectories (near the velar knee) displayed the greatest degree of error in the model. During speech production, for an oral to nasal production, the velum must move rapidly to make contact with the pharyngeal wall. This velocity and high deformability within short durations could attribute to the growth in error rate in the

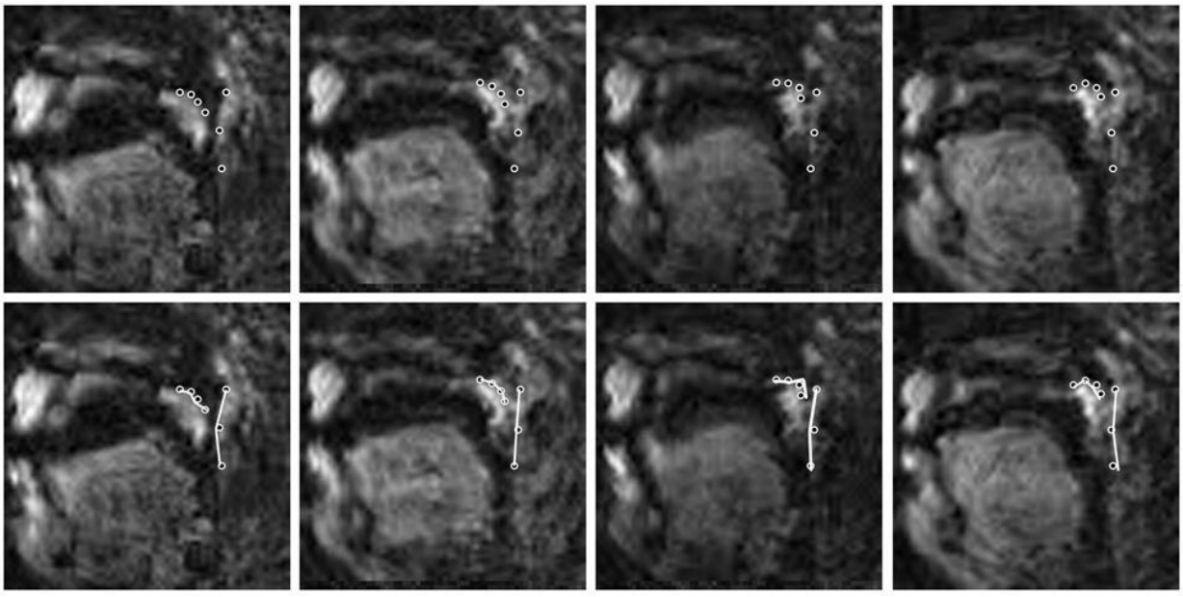

Figure 7: The top row represents the actual markers which are placed by researcher. The bottom row represents the superimposed actual markers and the predicted markers. In order to be distinguishable, predicted markers are connected.

prediction model. Because the speech sample contained more oral productions compared to nasal productions, the training protocol was dominant in velar positions during oral productions. It is possible that a more balanced speech production or speech stimuli containing no nasal sounds may produce a better model. The unbalanced training sets combined with the high level of deformability of the velum during nasal to oral productions likely contribute to the greater error rates noted along markers V3 and V4. A uniform sample distribution may solve this unbalanced training problem. A possible solution may be chunking audio signals into smaller segments which would contain uniform distribution. The input signal could thus be trimmed in which low and high amplitudes are balanced. Future studies could investigate the addition of a post processing phase and designing a smoother filter such as a Kalman filter [28] to make the prediction more accurate.

Image segmentation is used to cluster pixels into a designate regions using thresholding, histogram based, clustering, and edge detection methods. Edge detection is likely the most commonly used method in tracking the regions of the oral and pharyngeal structures. The primary disadvantage of this method, however, is the discontinuity that occurs when two structures of similar gradients communicate and the boundaries are lost. This is the primary concern in tracking velar movements due to the communication of the velum and posterior pharyngeal wall during closure. This method, therefore, cannot be successfully applied to velar tracking. The current model offers an advantage over this method in that it does not rely solely on edge tracking systems. HMM differs from other machine learning techniques; however, they have similar fundamental concepts. HMM requires 1) time series based images (image sequence) and 2) a synchronized associated audio. HMM is one of the most well-defined methods in computer speech recognition; however, it can be replaced with other prediction models. HMM is widely used in speech-to-text. The significant differences between the traditional application of HMM (speech-to-text) and the application of HMM in the present study are as follows:

1) The proposed model reconstructs location and figure of velum and pharyngeal wall for each frame, however in traditional speech-to-text approaches, a chunk of given signal is mapped to a word.

2) Compared to other classifiers, HMM displays a faster training and response time. This feature enables researchers to use this tool in real-time mode and monitor the structures on real-time, such as shown with dynamic MRI speech samples.

The primary advantage of this method is the ability to apply the trained model to a new set of patient data to predict velar positioning. This method is also advantageous because it overcomes the barriers commonly seen in edge detection methods when edges are difficult to delineate. Although error rate was relatively low, this is noted as the primary disadvantage and future developments should aim to decrease this error rate. While this error rate may be acceptable in normal anatomy, it may be problematic in disordered anatomy where velopharyngeal gap sizes may be minimal and thus not detected using these methods.

It was expected that the error rate would be the least for the nasal production compared to the sibilant and vowel production. This was expected due to the improved boundary created by the air interface while the velum is in the lowered position. When the velum is elevated, the boundary of the velum to the pharyngeal wall is difficult to separate. The sibilant /s/ production, however, was similar in error rate to that of the nasal production. Accumulative error and the audio signal amplitude showed a similar pattern in which the higher the amplitude generated in the signal, the higher the residual error that was introduced into the model.

Depending on the type of closure pattern, velopharyngeal movement typically involves superior and posterior movement of the velum, medial movement of the lateral pharyngeal walls, and anterior movement of the posterior pharyngeal wall. This sphincter-like action of the velopharyngeal port represents a three dimensional process. However, the present study focused solely on the two-dimensional process along the midsagittal plane. It demonstrated the success of a predictive model using audio signals to feed a computerized system to predict velopharyngeal structural positions in the midsagittal image plane. Future studies should aim to evaluate the system from an oblique coronal or axial image plane to evaluate if this model can predict velopharyngeal gaps and positions of the velum and pharyngeal wall from the portal view. As an experimental study in innovative methods, we did not include assess the application of this development in individuals with a velopharyngeal gap. We are currently working to enhance our model by decreasing the error rate so that methods are sensitive to small velopharyngeal gap sizes.

This study provides advancement in the area of dynamic MRI data processing of the velopharyngeal mechanism. Studies have demonstrated the potential benefits of using dynamic MRI for cleft palate speech assessments [14, 15, 17]. However, an obstacle for these studies has been the inability to rapidly produce meaningful clinical data to the practicing clinician and cleft palate team. Unlike traditional imaging methods (e.g., videofluoroscopy and nasoendoscopy), dynamic MRI requires post-processing of the data. Future studies should include larger sample sizes, subjects with abnormal velopharyngeal movements, and assess velopharyngeal closure from the plane in which closure occurs [14]. A child with velopharyngeal dysfunction (perceived hypernasality) may demonstrate a very small velopharyngeal gap. For this reason, the error threshold was set at 1 pixel.

Advantages of this approach include a novel method to evaluate the presence of a velopharyngeal gap using audio files. It is not expected that this innovation would replace any of the current instrumentation in cleft care which provides direct visualization of the velopharyngeal anatomy. Rather, this may be a cost-efficient method which may be particularly useful for school-based speech language pathologists who do not have access to traditional velopharyngeal instrumentation. In such, this tool may provide further support for the school-based clinician during the perceptual assessment of resonance. This may improve their confidence in making referrals for velopharyngeal dysfunction (as evidenced by perceptual speech assessment and a documented velopharyngeal gap) to a cleft palate craniofacial team for further evaluation. Disadvantages of the present study include the single study design and limiting interpretations to only the midsagittal image plane. Improvements in the model construction such as using a balanced (oral to nasal) training protocol may demonstrate an improved model.

CONCLUSION

This study demonstrates a potential method for using audio signals to determine velopharyngeal positioning during speech production. Although this study demonstrated a single case study, the findings illustrate a novel innovative model combining training and evaluating protocols that can be applied to any speech task. This is the first study to demonstrate automatic tracings of velar and pharyngeal movement along the midsagittal plane using dynamic MRI. The results from this study demonstrate a unique approach that relies on audio recordings of speech stimuli to provide a visual representation of velopharyngeal function. Future studies should investigate methods for velar prediction based on acoustic correlates in clinical populations such as cleft palate and dysarthria.

ACKNOWLEDGEMENTS

This study was made possible by grant number 1R03DC009676-01A1 from the National Institute of Deafness and Other Communicative Disorders. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Institutes of Health. The authors would like to thank Dr. Bradley Sutton and Dr. David Kuehn for their support in MRI methods and MRI data collection.

REFERENCES

[1] Hess U, Hanning C, Sader R, *et al*. Evaluation of velopharyngeal closure in preoperative planning of maxillary advancement. Rontegenpraxis 1996; 49: 25-26.

[2] Sader R, Horch HH, Herzog M, *et al*. High-frequency videocinematography for the objective imaging of the velopharyngeal closure mechanism in cleft palate patients. Fortschr Kieferorthop 1994; 55(4): 169-175. http://dx.doi.org/10.1007/BF02285407

[3] Skolnick ML, Cohn ER. Videofluoroscopic Studies of Speech in Patients with Cleft Palate. New York: Springer-Verlag 1989. http://dx.doi.org/10.1007/978-1-4613-8874-6

[4] Witzel MA, Stringer DA. Methods of assessing velopharyngeal function. Philadelphia: W. B. Saunders 1990.

[5] Pigott RW. An analysis of the strengths and weaknesses of endoscopic and radiological investigations of the velopharyngeal incompetence based on 20-year experience of simultaneous recording. Br J Plast Surg 2002; 55: 32-35. http://dx.doi.org/10.1054/bjps.2001.3732

[6] Pigott RW, Makepeace AP. Some characteristics of endoscopic and radiological systems used in elaboration of the diagnosis of velopharyngeal incompetence. Br J Plast Surg 1982; 35(1): 19-32. http://dx.doi.org/10.1016/0007-1226(82)90078-9

[7] Sinclair SW, Daviews DM, Bracka A. Comparative reliability of nasal pharyngoscopy and videofluorography in the assessment of velopharyngeal incompetence. Br J Plast Surg 1982; 35(2): 113-117. http://dx.doi.org/10.1016/0007-1226(82)90146-1

[8] Henningsson G, Isberg A. A cineradiographic study of velopharyngeal movements for deviant versus nondeviant articulation. Cleft Palate Craniofac J 1991; 28(1): 115-117. http://dx.doi.org/10.1597/1545-1569(1991)028<0115:ACSOVM>2.3.CO;2

[9] Karnell MP, Ibuki K, Morris HL, Van Demark DR. Reliability of the nasopharyngeal fiberscope (NPF) for assessing velopharyngeal function. Cleft Palate Craniofac J 1983; 20(3): 199-208.

[10] Birch MJ, Sommerland BC, Fenn C, Butterworth M. A study of the measurement errors associated with the analysis of velar movements assessed from lateral videofluoroscopic investigations. Cleft Palate Craniofac J 1999; 36(6): 499-507. http://dx.doi.org/10.1597/1545-1569(1999)036<0499:ASOTME>2.3.CO;2

[11] Havstam C, Lohmander A, Persson C, *et al*. Evaluation of VPI-assessment with videofluoroscopy and nasoendoscopy. Br J Plast Surg 2005; 58(7): 922-31. http://dx.doi.org/10.1016/j.bjps.2005.02.012

[12] Lam DJ, Starr JR, Perkins JA, *et al*. A comparison of nasoendoscopy and multiview videofluroscopy in assessing velopharyngeal insufficiency. Otolaryngol Head Neck Surg 2006; 134(3): 394-402. http://dx.doi.org/10.1016/j.otohns.2005.11.028

[13] Kuehn DP, Ettema SL, Goldwasser MS, Barkmeier JC, Wachtel JM. Magnetic resonance imaging in the evaluation of occult submucous cleft palate. Cleft Palate Craniofac J 2001; 38(5): 421-431. http://dx.doi.org/10.1597/1545-1569(2001)038<0421:MRIITE>2.0.CO;2

[14] Perry JL, Sutton BP, Kuehn DP, Gamage JK. Using MRI for assessing velopharyngeal structures and function. Cleft Palate Craniofac J 2014; 51(4): 476-485. http://dx.doi.org/10.1597/12-083

[15] Sutton BP, Conway CA, Bae Y, Seethamraju R, Kuehn DP. Faster dynamic imaging of speech with field inhomogeneity correlated spiral fast low angle shot (FLASH) at 3T. J Magn Reson Imaging 2010; 32(5): 1228-1237. http://dx.doi.org/10.1002/jmri.22369

[16] Fu M, Bo Z, Shosted RK, *et al*. High-resolution dynamic speech imaging with joint low-rank and sparsity constraints. Magn Reson Med 2015; 73(5): 1820-1832. http://dx.doi.org/10.1002/mrm.25302

[17] Perry JL, Kuehn DP, Sutton BP, Fang X. Analyses of velopharyngeal function in children using real-time dynamic MRI. Cleft Palate Craniofac J; in press

[18] Bae Y, Kuehn DP, Conway CA, Sutton BP. Real-time magnetic resonance imaging of velopharyngeal activities with simultaneous speech recordings. Cleft Palate Craniofac J 2011; 48(6): 695-707. http://dx.doi.org/10.1597/09-158

[19] Jelinek F. Statistical methods for speech recognition. Massachusetts: MIT press 1998.

[20] Li Q, Soong FK, Siohan O. A high-performance auditory feature for robust speech recognition. In Proc. of the 6th International Conference on Spoken Language Processing (ICSLP) 2000; pp. 51-54.

[21] Han W, Chan CF, Choy CS, Pun KP. An efficient MFCC extraction method in speech recognition. In Proc. of the IEEE International Symposium on Circuits and Systems (ISCAS) 2006; pp. 145-148.

[22] Ghitza O. Auditory models and human performance in tasks related to speech coding and speech recognition. Speech and Audio Processing IEEE Proceedings 1994; 2(1): 115-132. http://dx.doi.org/10.1109/89.260357

[23] Sutton BP, Conway C, Bae Y, Brinegar C, Liang ZP, Kuehn DP. Dynamic imaging of speech and swallowing with MRI. In Proc. of IEEE Eng Med Biol Soc 2009; pp. 6651-6654. http://dx.doi.org/10.1109/iembs.2009.5332869

[24] Hodgson J. Understanding Records. 1st ed. London: Bloomsbury Academic 2010.

[25] Slaney M. Auditory toolbox [Internet]. Interval Research Corporation [cited 2015 November 8]. Available from: https://engineering.purdue.edu/~malcolm/interval/1998-010/

[26] Rabiner L. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 1989; 77(2): 257-286. http://dx.doi.org/10.1109/5.18626

[27] Forney GD. The Viterbi algorithm. Proceedings of the IEEE 1973; 61(3): 268-278. http://dx.doi.org/10.1109/PROC.1973.9030

[28] Welch G, Bishop G. An Introduction to the Kalman filter [Internet]. Department of Computer Science, University of North Carolina at Chapel Hill [cited 2015 November 20]. Available from: http://www.cs.unc.edu/~tracker/media/pdf/SIGGRAPH2001_CoursePack_08.pdf

Received on 01-03-2016 Accepted on 22-03-2016 Published on 14-05-2016

DOI: http://dx.doi.org/10.12970/2311-1917.2016.04.01.1