From experiments to articulatory motion—A three dimensional talking head model

Xiao Bo Lu 1, William Thorpe 1, Kylie Foster 2 and Peter Hunter 1

1 Bioengineering Institute, the University of Auckland, Auckland, New Zealand
2 Department of Food and Health, the University of Massey, Auckland, New Zealand
xiao.lu@auckland.ac.nz

1. Abstract

The goal of this study is to develop a customised computer model that can accurately represent the motion of vocal articulators during vowels and consonants. Models of the articulators were constructed as Finite Element (FE) meshes based on digitised high-resolution MRI (Magnetic Resonance Imaging) scans obtained during quiet breathing. Articulatory kinematics during speaking were obtained by EMA (Electromagnetic Articulography) and video of the face. The movement information thus acquired was applied to the FE model to provide jaw motion, modeled as a rigid body, and tongue, cheek and lip movements modeled with a free-form deformation technique. The motion of the epiglottis has also been considered in the model.

2. Introduction

Over the years, people have tried to build detailed models of the human vocal system to study human speech physiology. Modern imaging techniques such as MRI have allowed the creation of highly detailed and anatomically accurate articulator models. However, due to the slow imaging acquisition rate of MRI (if high resolution is to be obtained), such models are constrained to static configurations of the articulators. Articulatory dynamics can be measured by techniques such as EMA which track the 3-D movements of discrete points during speech utterances. Here we combine the high resolution of MRI with the fast sampling rate of EMA and video acquisition to create a realistic 3-D dynamic model of the human vocal system.

3. Static model construction

The MRI scans were conducted in a 1.5T Siemens Magnetom Avanto MRI scanner located at the Centre for Advanced MRI of the University of Auckland. The subject was a 1.83m tall 43 year old male native speaker of New Zealand English. Using a T2-weighted image, 585ms echo time and 3200ms repetition time, a stack of 160 256x256mm sagittal images of 1mm isotropic resolution was collected while the subject was lying supine and breathing quietly with mouth closed. A few sustained vowels and consonants were also scanned using a 20-slice fast scanning sequence to provide reference configurations for different speech sounds [1].

The acquired images during the quiet breathing stage were imported into the Zinc digitiser (ftp://ftp.bioeng.auckland.ac.nz/cmsis/mozcmgui/) for the segmentation. As illustrated in Figure 1, the stack of parallel sagittal images were loaded into a 256x160x256mm volume texture. From this volume texture, axial and coronal image slices were extracted to complement the original sagittal images. Boundaries of all the articulators were manually digitised by marking points along the tissue boundaries on each slice, spaced approximately 3mm apart. Most of the structures (i.e., tongue, cheeks, lips, larynx, epiglottis, jaw, maxilla, pharyngeal wall and upper face) were digitised on the axial images, with references to the intersecting sagittal slices to resolve ambiguities. We did not include the nasal cavity into the model at this stage, so the soft palate had to be in an elevated position, but it was in a descended position in the quiet breathing scans. Hence the shape of the soft palate was extracted from the coronal slices collected during one of the sustained vowels.

Figure 1: Digitisation of the tongue in the Zinc digitiser.

Figure 2: The subject-specific finite element vocal articulator models.

The resulting data clouds represent the exact physical dimensions of each object. These data clouds were then used to construct FE meshes where either the surface or volume is explicitly specified in 3-D space by their nodal values and interpolation schemes. Among different models, the maxilla (upper jaw),
jaw, tongue, cheeks and lips (together in one single mesh), soft palate and epiglottis are built as volume meshes, while the upper face and pharyngeal wall are made of only surface elements. All these geometric models were initially hand-drafted in Cmgui (www.cmiss.org/cmgui), then fitted to their corresponding data clouds via a least-square fitting algorithm [2] designed to minimise the projection distances from the data points to the mathematical surfaces through nodal parameters. Cubic hermite elements and Sobelov smoothing factors were used to remove any unrealistic edges on the models. The final assembly of the ‘Talking Head’ model consists of nine components as shown in Figure 2 and an overview of the whole model are given in Figure 3a and Figure 3b.

To facilitate tracking, 5mm circular reflective spots were put onto various facial points as visual markers (see Figure 4b). The remaining six EMA channels were all placed onto the tongue surface as shown in Figure 4b.

4. EMA-Video experiment

EMA is a technique which has been widely used in studies of articulatory motions of human speech. It involves the use of tri-axial alternating magnetic fields and miniature inductive sensors glued to the articulators to trace their movements. Among other studies using EMA to reconstruct articulator geometries, Kaburagi et al. [4] demonstrated a method to derive the shape of the mid-saggital contour from the positions of EMA points on the tongue surface and produced an estimation error within 1mm. Also, Engwall [5] in 2003 presented a 3D tongue model which utilised EMA data to control its motion.

The Carstens Articulograph AG500 was used for the experiment. The machine has 12 sensors, each capable of recording 3-D position and 2 rotation angles at 200Hz, and an analog audio channel recording at 16kHz. It has a measurement accuracy of 0.5mm [6]. In addition, we recorded video of the subject’s face from both frontal and left profile views using two Sony DCR-SX40 Digital hand cameras, at 30 frames per second. The setup of the experiment is given in Figure 4a.

To facilitate tracking, 5mm circular reflective spots were put onto various facial points as visual markers (see Figure 4b). Three reference EMA sensors were glued onto one of the upper incisors, and over the temporal bones beneath the right and left ears, respectively. The jaw movements were traced by the two sensors on the two lower canines. A sensor glued in the middle of the upper lip was also co-registered with a visual marker. The remaining six EMA channels were all placed onto the tongue surface as shown in Figure 4b.

The same subject as the one in the MRI scans participated this EMA-Video experiment. The data consists of a selection of short English words and sentences focusing on a range of vowels, stops, fricatives, nasals, liquids and consonant clusters. They were chosen so that:

1) All phonemes are in initial, medial and final positions;
2) Combinations of initial, medial and final consonants were paired with each of the selected vowels;
3) Consonant cluster were mixed with a wide range of vowels.

Each word and sentence were recorded in separated audio and EMA files and repeated twice consecutively. In total, there were 208 words and 13 sentences being recorded.

4.1. EMA data processing and registration

The positions of the EMA sensor within the system were optimised by the Carstens Calcpos program, based on the signals collected at 6 receivers. After the calculation, the calculated 3-D coordinates were normalised with respect to the head movements. Three of the sensors (left and right ears and upper incisor) served to create a head axis used to compensate for the head motion, using the Carstens NormPos program. The original EMA signal was sampled at 200 Hz but was resampled to 30Hz by custom programs written in Matlab (Mathworks) that matched the frame rate of the video data.

The data acquired by the EMA sensors were obtained in the local coordinate system of the machine. To apply them to our computer models, we transformed them into the modelling coordinate system. Both coordinate systems are of rectangular Cartesian type, so the task was to overlay the origins and align the axes. This was done by exporting the reference EMA frame measured at quiet breathing state into a visualisation package (Cmgui) together with the static models. Since the static models were also created in the rest breathing state of the same subject, we expect these sensors to be on the exact or very close anatomical locations in the model as where they were glued onto the subject in the EMA experiment. The EMA frame was manually rotated and translated to match their target locations and the resulting transformation matrix was recorded, and then applied to the remainder of the EMA data to transform them into the modelling coordinate system.

4.2. Video segmentation

Two sets of video images (front view and side view) were used to provide information on the lips area. The video frames were sampled at about 30Hz - the same as the resampled EMA data. The first task was to synchronise the images with the EMA data. This was done with the aid of the audio files from both machines. Spectrograms of the audio files were used to precisely match the onset of the speech signals in each recording and thus temporally align the video and EMA data.
The shape of the lip was manually measured image by image. It was assumed the front view aligned with the y-z plane on the model and the side view aligned with the x-z plane. Relative positions were measured between every visual marker and the reference marker (in the middle of the upper lip co-registered with an EMA sensor). The y and z coordinates were measured on the front views while the side view provided the x coordinates. With these spatial information and symmetry assumptions, plus the EMA measurements of the lip sensor, the 3-D locations of the visual markers could be fully reproduced in the modelling frame.

5. Simulating articulatory motions

There are four dynamic articulators included in the model – the jaw, tongue, cheek/lips and epiglottis. Their motion are simulated separately using different methods, with connections to link their relative movements against each other.

5.1. The jaw kinematics

In the current model, the jaw is set to perform rigid body motion (i.e., rotation and translation) during all the speech utterances. The Carsten AG500 system provides both positional and rotational information of each sensor. As a result, we can calculate 4 sets of coordinates from the two jaw sensors at each sampling instant. We needed to derive Euclidean transformation tensors from these coordinates. First we calculated the affine transformation tensor defined as:

\[ T_{affine} = \begin{bmatrix} t_{11} & t_{12} & t_{13} & t_{14} \\ t_{21} & t_{22} & t_{23} & t_{24} \\ t_{31} & t_{32} & t_{33} & t_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix} \]

(1)

which can be calculated from the four pairs of deformed \((x, y, z)\) and undeformed coordinates \((X, Y, Z)\), using:

\[
\begin{bmatrix} X_1 & Y_1 & Z_1 & 1 \\ X_2 & Y_2 & Z_2 & 1 \\ X_3 & Y_3 & Z_3 & 1 \\ X_4 & Y_4 & Z_4 & 1 \end{bmatrix} \begin{bmatrix} t_{11} \\ t_{12} \\ t_{13} \\ t_{14} \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}
\]

(2)

The resulting transformation includes rotation, translation, shearing and scaling. To get rid of the shearing and scaling components (we assume these should be negligible based on constraints of how the jaw can move, so any such components probably result from noise in the measurements), we made use of the orthogonality property of the rotational matrix. First we singled out the inner 3 \(\times\) 3 matrix which contains the rotation, shearing and scaling components. The resulting matrix can be written as:

\[ T_{3x3} = RU \text{ and } T_{3x3}^T T_{3x3} = U^T R^T RU = U^2 \]

(3)

where \(R\) is the rotational tensor and is an orthogonal matrix, and \(U\) is the matrix containing the shearing and scaling components. Then the scaling and shearing parts of the transformation can be stripped off via:

\[ R = T_{3x3} (T_{3x3}^T T_{3x3})^{-\frac{1}{2}} \]

(4)

Finally, we substitute the inner 3 \(\times\) 3 matrix of \(T\) with the rotational tensor \(R\) to get the Euclidean transformation for the jaw. An example video of the simulated jaw motion during the word ‘hoard’ is provided in the file jaw.mpg.

5.2. Deforming the tongue, cheeks and lips

The host mesh fitting technique described in [3] is designed for an under-defined problem where the information we have is not sufficient to provide a unique solution. It can be broken down into three basic steps: 1) Reduce the number of DOFs (Degrees of Freedom) in the problem; 2) Specify the known DOFs; 3) Provide guesses for the still undefined DOFs. Here, this allowed us to transform complex structures (the tongue and lips) based on the movements of a few anatomical points on the objects’ surfaces.

The first step is embedding the complex object (the slave mesh) into a simple object with fewer DOFs (the host mesh). In the case of the tongue, (Figure 5a), we embedded the tongue mesh (36 cubic hermite nodes and 864 DOFs) into a mesh made of a single cubic element (8 cubic nodes, 192 DOFs). On the cheek and lips, the original mesh contains 210 cubic hermite nodes (5052 DOFs) and has been placed into a host of 75 cubic hermite nodes with fixed (1575 DOFs) nodal coordinates being (see Figure 5c).

In the second step, objective functions are set up to deform the host mesh according to the movements of a few landmark points defined on the slave object’s surface. By doing so, the embedded structure is also deformed as part of the host as illustrated in Figure 5b and Figure 5d. On the tongue, the landmark points were set to the EMA sensor positions, with their new locations fed into the host-mesh fitting algorithm at each time step. There are only 6 EMA sensors on the tongue so there are only 18 DOFs defining the transformation. The landmark points used in deforming the cheek and lips mesh were collected from the visual markers (12 points, 36 DOFs).

It is also possible to introduce other physiological constraints on the host mesh, including fixed boundary positions such as points where the deformable object joins rigid articulators. Since the tongue is mechanically coupled to the jaw mainly through the muscle layers and connective tissues attached to the tongue root, another 8 landmark points were defined halfway around the tongue body (see Figure 5a), with their movements tied to the jaw motion at each time step. In the case of cheek and lips, additional ‘attachment points’ were also specified on the inner surface and linked to the jaw motion.

The problem is still under-defined (e.g., 192 DOFs in the tongue host vs 42 specified DOFs) so in the third step, the solution is constrained by a penalty function [3] made of Sobolev smoothing weights introduced to constrain the geometric features of the host mesh (e.g., arc length, curvature, face area and volume). The chosen penalty values are then optimised manually so that the solution provides physiologically realistic deformations. More details of the modelled tongue and cheek/lips motion during the word ‘hoard’ can be viewed in the video files tongue.mpg and cheek_lips.mpg.

5.3. Tongue-Epiglottis contact

We do not have any direct EMA measurements in this region, however we do know the epiglottis moves as the tongue shifts towards the back like the situations in back vowels. We therefore assume the motion of the epiglottis is only a passive reaction to the contact by the tongue. Furthermore, the epiglottis is treated as a rigid body and rotates about the point where it is attached to the tongue root and hyoid bone. An algorithm was
The tongue and its host mesh in the resting state.

The tongue and its host mesh in deformed state during the vowel /ɔ/. 

The cheek/lips and the host mesh in resting state.

The cheek/lips and the host mesh in deformed state during the vowel /ɔ/. 

Figure 5: The host mesh fitting for the tongue and cheek/lips. The landmark points are shown by the green spheres.

written based on these assumptions to model the epiglottis motion. It can be summarised into two stages: collision detection and collision correction. The former is achieved by projecting points sitting on the back surface of the epiglottis onto the frontal surface of the epiglottis and the surface of tongue. A collision occurs when the projection distance to the tongue is less than the respective distance to the frontal surface of the epiglottis. When a collision is detected, the epiglottis is rotated about its root until the collision detection routine returns a false. A generic example is demonstrated in Figure 6 and the video clip (epiglottis.mpg).

(a) Before contact.
(b) Contact 1.
(c) Contact 2.

Figure 6: Contact simulation between the tongue and epiglottis.

6. Results

Figure 7 illustrates the simulated stance for the vowel /ɔ/ during the word ‘hoard’. The word lasted about 1 second in real time and constitutes 28 frames in the sampled data (EMA and video). Each frame took 68 seconds simulation time on a single intel core processor. As we can see in Figure 7, the talking head model could in general reproduce the articulatory configuration for the vowel in the mouth region. The lips are more protruded in the simulated frame than the imaged ones, which could be explained by the effect of artificial sustaining during the MRI scan. On the other hand, the tongue is not as low as it should be.

This could be mainly due to the fact that there is no kinematic information available at the back of the tongue. There are also some interceptions among different articulator meshes, but it does not affect the geometry of the vocal tract.

(a) The mid-sagittal view of simulated vowel.
(b) The mid-sagittal view of sustained vowel by MRI.

Figure 7: The vowel /ɔ/ in word ‘hoard’.

7. Conclusion

The study shows that it is possible to use measurements of a few discrete points to control the motion of a fully 3-D articulatory model during normal speech. Such a model forms the basis for a complete 3-D articulatory speech synthesis which can help us to study linkages between the acoustic features of the voice and their physiological cues. So far, only the EMA and video data have been incorporated into the model to provide articulator controls. In the future, more techniques like CT (computed tomography), fast MRI and ultrasound could be utilised to complement the existing experiment tools and provide kinematic information in the larynx and pharynx regions.

8. Acknowledgements

The authors thank Anna-Maria Lydon for her assistance with the MRI scans and Margaret Maclagan for the phonetically balanced list of words used in the EMA experiment.

9. References