SUBJECT-SPECIFIC BIOMECHANICAL MODELLING OF THE OROPHARYNX WITH APPLICATION TO SPEECH PRODUCTION

N. M. Harandi [⋆†], *J. Woo* [††], *M. R. Farazi* [⋆†], *I. Stavness* [⋆⋆], *M. Stone* [°], *S. Fels* [†], *R. Abugharbieh* [⋆]

[⋆]BiSICL and [†]HCT Lab, Electrical Engineering Department,University of British Columbia, Canada
[⋆⋆]Department of Computer Science, University of Saskatchewan, Canada
[††]Department of Radiology, Harvard Medical School, USA
[°]Department of Orthodontics, University of Maryland School of Dentistry, USA

ABSTRACT

In this work, we develop a 3D subject-specific biomechanical model of the oropharynx in order to investigate and simulate speech production. Our muscle-activated model is generated based on the subject-specific anatomy captured from dynamic volumetric cine-MRI data. Our model includes an air-tight deformable airway that enables speech synthesis. We simulate our model based on actual tissue motion tracked from the tongue during speech production, which we extract from the tagged-MRI data. We quantitatively validate our model on MRI data achieving an average target point tracking error of $1.15mm \pm 0.632$, and an acoustic formant frequency estimation error of $6.01\% \pm 4.92\%$.

Index Terms— subject-specific modelling, inverse simulation, oropharynx, speech production, cine-MRI

1. INTRODUCTION

Speech production involves synchronized motion of the oropharyngeal structures initiated by a complex set of neural excitations of the corresponding muscles. Speech impediments are widespread and result from complicated mental or physiological disorders. Understanding the exact mechanism of speech – including the biomechanics and motor control – is beneficial for addressing the cause of the impediment and planning an effective treatment for the patient.

Physics-based modelling of the oropharyngeal structures has been reported in the context of studying speech motor control [1]. A genetic coupled biomechanical model of the tongue-jaw-hyoid has been implemented in the ArtiSynth simulation framework (www.artisynth.org)[2]. Further, the articulators were enclosed with a deformable model of the airway to enable articulatory speech synthesis [3] (see figure 1). However, current biomechanical models of oropharynx remain generic and do not provide individualized information. Real-time medical imaging technologies have provided the possibility to capture subject-specific dynamic physiology of speech. For example, dynamic MRI is able to capture soft-tissue articulators – the tongue, soft palate, epiglottis and lips – during consecutive repetitions of an speech utterance. Tagged-MRI, in particular, was used to compute the displacement field of tongue tissue points in high resolution [4]. Such data renders adaptation of generic models to fit the subject domain feasible, which in turn facilitates the investigation of the inter and intra-subject variability of speech.

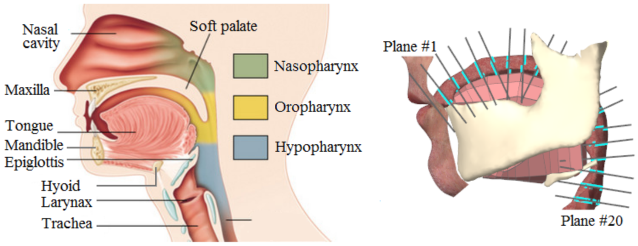

Fig. 1: Head and neck: Anatomy (left) vs. generic biomechanical model in ArtiSynth (right). Note that the planes are orthogonal to the vocal tract center line and evenly distributed from the lips (#1) to below the epiglottis (#20).

Forward dynamic simulation requires fine tuning of muscle activations over time. Electromyography (EMG) recordings of speech have been previously considered but lack a suitable technology to deal with the moist surface and highly deformable body of the tongue [5]. Furthermore, the relationship between EMG signals and muscle forces is not straight forward. As an alternative, muscle activations can be predicted from kinematics by solving an inverse problem [2].

We investigate subject-specific articulatory synthesis of speech based on biomechanical simulation of the oropharyx in ArtiSynth. In an earlier work, we developed a work-flow for subject-specific modelling and simulation of the tongue according to MRI data [6]. In this paper, we create subject-specific models of the oropharyngeal bones and enable speech synthesis by adding a deformable air-tight model of the vocal tract. We couple our models together, accounting for the contact force between the maxilla and the tongue in case of collision. Finally, our biomechanically-deformed model of vocal tract enables us to use our model in conjunction with an articulatory acoustic synthesiser [3, 7].

2. MATERIALS AND METHODS

Figure 2 shows the modular architecture of our proposed work-flow for subject-specific simulation and synthesis of speech. After cine and tagged-MRI are acquired during repetitions of the speech utterance, we register our previously developed generic coupled biomechanical model of the tongue-jaw-hyoid [2] to fit our subject's geometry as captured in the first time-frame of the cine-MRI data. We then simulate the speech kinematics by solving an inverse problem for muscle activations, given the trajectory of the tongue tissue points that we extract from tagged-MRI. In order to evaluate the acoustic functionality of our subject-specific model, we solve a 1D implementation of the Navier-Stokes equation for the deformed shape of our vocal tract [7].

2.1. Data Acquisition and Processing

Our MRI data captures a 22-year-old American white male with mid-Atlantic dialect repeating the utterance *a-geese* (/ə-gis/) to the metronome. Both tagged and cine-MRI were acquired with 1.875 mm in-plane (dense) and 6.00 mm slice resolution (sparse). Super resolution MRI volumes were reconstructed with an isotropic resolution of 1.875 mm, for each of the 26 time frames [4].

Tissue points of the tongue were tracked by using both the estimated motion from tagged-MRI and the surface information from cine-MRI. A 3D dense and incompressible deformation field was reconstructed from tagged-MRI based on the harmonic phase algorithm [8]. The 3D deformation of the surface was computed using diffeomorphic demons in cine-MRI [9]. The two were combined to obtain a reliable displacement field as described in [4]. To reduce the noise in the spatial domain, we average the displacements vectors in the neighbourhood of each target point. Also, to create a smooth motion over time, we select 6 main key time-frames of our speech utterance; we then perform a cubic interpolation to calculate the displacement field at the intermediate time-frames.

2.2. Oropharyngeal Model

The generic FE tongue model available in ArtiSynth provides 2493 DOFs and consists of 11 bilateral muscle groups [1] as detailed in [10]. We use the Blemker muscle model with Mooney-Rivlin material to ensure hyper-elasticity, and consider the effect of passive muscle forces [11]. This model is coupled with the jaw and hyoid rigid bodies via multiple attachment points [12]. To create the subject-specific model, we first delineate the surface geometry of the tongue in the first time-frame of the cine-MRI volume using *TurtleSeg*, a semi-automatic segmentation tool suitable for delineation of low-contrast soft-tissues [13]. We use a multi-scale, iterative and elastic registration method called Mesh-Match-and-Repair to fit the generic volumetric FE to the segmented subject surface mesh [14]. The registration starts by matching the two surfaces, followed by the application of the computed deformation field to the inner nodes of the FE model via interpolation. A follow-up repair step ensures all elements in the FE model satisfy the minimum quality standard specified by ANSYS simulation framework (www.ansys.com).

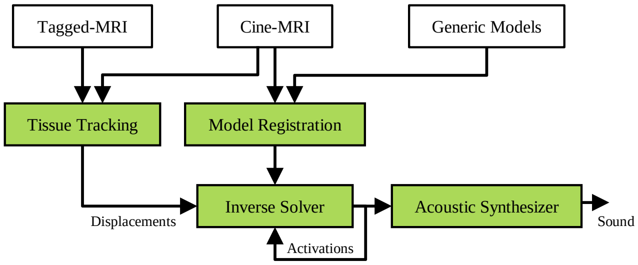

Fig. 2: Proposed work-flow for subject-specific biomechanical speech simulation.

Our proposed approach deploys the generic jaw-hyoid model developed in ArtiSynth [12], which includes rigid-bodies for the mandible and hyoid, as well as 13 pairs of bilateral point-to-point Hill-type actuators [2]. This generic model also provides surface constraints for the temporomandibular joints. To create the similar model for our subject, we segment the bone structures from the first time-frame of cine-MRI. However, since bone is partially visible in MRI, the result surfaces are not anatomically complete nor of sufficient mesh quality. To address this issue, we register the generic models of the mandible and hyoid to our partially-segmented surfaces using the coherent point drift (CPD) algorithm [15]. CPD is robust in the presence of outliers as well as missing points, and results in smooth meshes that now include all the important anatomical landmarks.

In order to model the vocal tract, we use a deformable skin which is set initially to match the geometry of the airway in the first time-frame of the cine-MRI. The skin is attached to and deforms along with the motion of the tongue and mandible; we also restrict the deformations to the the fixed boundaries of the maxilla and pharyngeal wall. The position of each skin vertex, $\mathbf{q_v}$, is calculated as a weighted sum of contributions from its master component (e.g. FE nodes of the tongue or rigid body frame of the jaw and mandible):

$$\mathbf{q}_v = \mathbf{q}_{v_0} + \sum_{i=1}^{M} w_i f_i(\mathbf{q}_m, \mathbf{q}_{m_0}, \mathbf{q}_{v_0}) \qquad (1)$$

[1] Genioglossus: anterior (GGA), medium (GGM), posterior (GGP); hyoglossus(HG); styloglossus (STY); inferior longitudinal(IL); verticalis (VERT); transverses (TRANS); geniohyoid (GH); mylohyoid (MH); superior longitudinal(SL).

[2] Mylohyoid: anterior(AM), posterior (PM); temporal: anterior (AT), middle (MT), posterior (PT); masseter: superficial (SM), Deep (DM); pterygoid: medial (MP), superior-lateral (SP), inferior-lateral (IP); digastric: anterior (AD), posterior (PD); stylo-hyoid(SH).

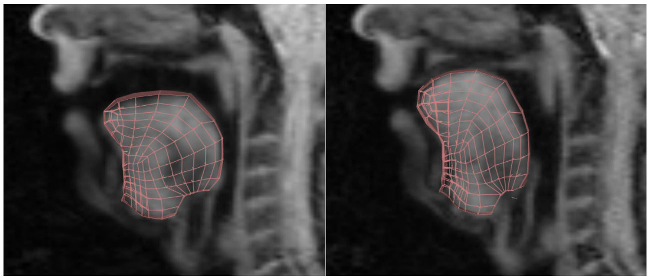

Fig. 3: Midsagittal slice of the subject-specific FE tongue model overlayed on the cine-MR images in the 5th (left) and 17th time-frames (right) corresponding to the (/ə/) and (/i/) respectively.

where $\mathbf{q}_{v_0}$ is the initial position of the skinned point, $\mathbf{q}_{m_0}$ is the collective rest state of the masters, w_i is the skinning weight associated with the ith master component, and f_i is the corresponding blending function as described in [16]. To provide two-way coupling , we propagate forces acting on the skin points back to their dynamic masters.

2.3. Inverse Simulation

Forward-dynamics simulation in ArtiSynth computes the system velocities, $\mathbf{u}$, in response to muscle-activation-dependent forces. Inverse simulation, one the other hand, estimates the muscle activations that yield a given set of target velocities $\mathbf{v}$, defined in a sub-space of $\mathbf{u}$. The inverse solver computes the normalized activations $\mathbf{a}$, by solving a quadratic program subject to the condition $0 < \mathbf{a} < 1$:

$$\mathbf{a} = argmin(\|(\mathbf{v} - \mathbf{Ha})\|^2 + \alpha\|\mathbf{a}\|^2 + \beta\|\dot{\mathbf{a}}\|^2) \qquad (2)$$

where $||\mathbf{x}||$ and $\dot{\mathbf{x}}$ denote the norm and time-derivative of $\mathbf{x}$; $\mathbf{H}$ is a matrix that summarizes the biomechanical characteristics of the system such as mass, joint constraints and force-activation properties of the muscles; α and β are ℓ^2-regularization and damping coefficients. The estimated muscle activations are fed back to the forward dynamics system to provide extra feedback to the inverse solver. The solution converges after limited number of iterations [2]. We set $\alpha = \beta = 0.005$ and use 21 target points in the left half of the tongue, while enabling bilaterally symmetric muscle excitation. Our proposed distribution of the target points provides adequate tracking information for each individual muscle exciter, and does not overly constrain a single element.

2.4. Acoustic Synthesis

Articulatory speech synthesizers generate sound based on the biomechanics of speech in the upper airway. Vibration of the vocal folds under the expiatory pressure of the lungs acts as the input to the system. The vocal tract constitutes a filter where sound frequencies are shaped. This creates a number of resonant peaks in the spectrum, known as formants. The first and second formants – F_1 and F_2 – are mainly used to define distinct phonemes of speech. The value of F_1 and F_2 depends on the height and backness-frontness of the tongue body respectively.

Table 1: Summary of the active muscles during simulation of the speech utterance *a-geese*.

Phoneme	Tongue Muscles	Jaw Muscles
/ə/	GGA, GGM, HG	IP
/g/	GGP, STY, TRANS, SL, MH	IP, SM
/i/	GGP, VERT, TRANS, STY	SM, AT, PD, MT, MP
/s/	MH, GH, GGM, SL, TRANS	–

In our model, we manipulate the shape of the vocal tract using the muscle activations computed from the inverse simulation. The resonating tube is represented as a transfer function defined by the cross-sectional areas of 20 segments along the vocal tract. The glottal sound source is a two-mass model of the oscillations of the vocal folds. We couple the source to the filter and solve a 1D implementation of the Navier-Stokes equations as described in [7].

The vocal folds oscillate during the vowels and voiced constants (such as /m/ or /n/), but are widely open and of minimal effect in fricatives (such as /s/). Constrictions at certain points in the tract create turbulence that generates high frequency noise responsible for making the fricatives. The synthesis of fricatives depends highly on lung pressure and noise characteristics of the system. Due to the lack of such information, we solely focus on synthesis of vowels /ə/ and /i/ in the utterance *a-geese* which correspond to the time-frames 5 and 17 in the cine-MRI.

3. RESULTS AND DISCUSSION

Figure 3 shows the deformed shape of our subject-specific tongue model – for the vowels /ə/ and /i/ – with respect to the midsagittal cine-MRI. Our average tracking error – defined as the distance between the position of the target points in our simulation and in the tagged-MRI – was 1.15mm $\pm$ 0.632, which is within the accuracy range of the tagged-MRI and sufficient for speech simulation purposes.

Table 1 provides a summary of active muscles for each speech phoneme in the utterance *a-geese*. To move from a rest position in the front of the mouth to /ə/, the tongue lowers and retrudes, using the GGA, GGM and HG. Motion into /g/ requires upward motion of the entire tongue, engaging the GGP, STY, TRANS, SL and MH. Subsequent forward motion into /i/ would further engage the GGP, VERT, SL and TRANS. Motion into /s/ would engage the MH, GH, GGM, SL and TRANS. The majority of jaw muscles activate throughout /i/ and end after /s/ begins. The IP exhibits activation pulses which protrude the jaw slightly to start /g/ and considerably more for /i/. SM – which elevates the mandible – activates during /g/ and more so during /i/.

In addition, we manually segmented the vocal tract from time-frames 5 and 17 of the cine-MRI and compared their

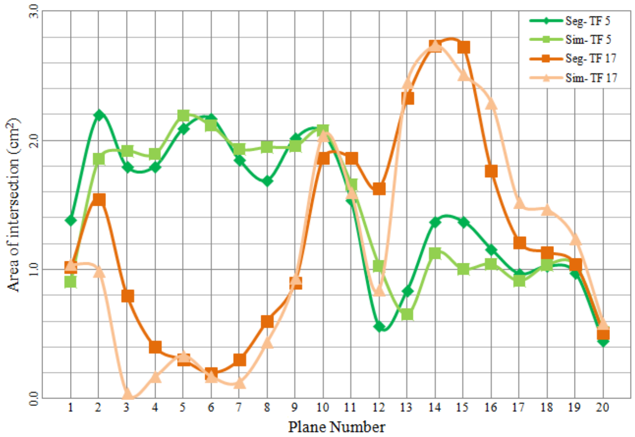

Fig. 4: Cross-sectional area profile for simulation (Sim) vs. ground truth (Seg) at time-frames (TF) 5 and 17 of the cine-MRI.

corresponding area profiles with the result of our subject-specific simulation (see figure 4). Note how our model is able to capture the expected shape of the vocal tract. The largest mismatch happens at plane #1 for /ə/ and at plane #12 for /i/, which in fact corresponds to the lips and the soft palate which were not included in our model. Finally, we calculate the (F_1,F_2) formant frequencies for our simulation, time-frame 5: (556Hz,1235Hz) and time-frame 17: (233Hz,1845Hz) and compare to the ground truth values calculated from cine-MRI at time-frame 5: (571Hz,1312Hz) and time-frame 17: (268Hz,1896Hz). This corresponds to $6.01\% \pm 4.92\%$ average estimation error in the calculated values of the two first formants.

4. CONCLUSION

We proposed an approach for subject-specific modelling and simulation of the oropharynx to enable speech synthesis. Our model is able to follow the deformation of the tongue tissue in tagged-MRI data, estimating plausible muscle activations, along with acceptable acoustic responses. The modular architecture of our framework can benefit from further improvements in each individual modules such as a higher-resolution generic tongue model and a more advanced speech synthesizer. In the future, we plan to adapt the generic ArtiSynth models of the lips, soft palate and epiglottis into our subject-specific tracking and simulation platform. We also plan to explore inter-subject variability of speakers by expanding our experiments to include more male and female subjects.

5. ACKNOWLEDGEMENTS

This work is funded by Natural Sciences and Engineering Research Council of Canada (NSERC), NSERC-Collaborative Health Research Project (CHRP) and Network Centre of Excellence on Graphics, Animation and New Media (GRAND).

6. REFERENCES

[1] J. M. Gérard et al. “A 3D dynamical biomechanical tongue model to study speech motor control”. In: *arXiv preprint physics/0606148* (2003).

[2] I. Stavness et al. “Automatic prediction of tongue muscle activations using a finite element model”. In: *J Biomech* 45 (2012), pp. 2841–2848.

[3] K. van den Doe et al. “Towards articulatory speech synthesis with a dynamic 3D finite element tongue model”. In: *Proc Int Semin Speech Prod (ISSP)*. 2006, pp. 59–66.

[4] F. Xing et al. “3d tongue motion from tagged and cine MR images”. In: *Proc Med Image Comput Comput Assist Interv (MICCAI)*. 2013, pp. 41–48.

[5] K. Yoshida et al. “Clinical science EMG approach to assessing tongue activity using miniature surface electrodes”. In: *J Dent Res* 61 (1982), pp. 1148–1152.

[6] N. M. Harandi et al. “Subject-specific biomechanical modelling of the tongue: analysis of muscle activations during speech”. In: *Proc Int Semin Speech Prod (ISSP)*. 2014, pp. 174–177.

[7] K. van den Doel and U. M. Ascher. “Real-time numerical solution of Webster’s equation on a non-uniform grid”. In: *IEEE Trans Speech Audio Process* 16 (2008), pp. 1163–1172.

[8] N. F. Osman et al. “Imaging heart motion using harmonic phase MRI”. In: *IEEE Trans Me. Imag* 19 (2000), pp. 186–202.

[9] T. Vercauteren et al. “Diffeomorphic demons: efficient non-parametric image registration”. In: *NeuroImage* 45 (2008), pp. 61–72.

[10] S. Buchaillard et al. “A biomechanical model of cardinal vowel production: Muscle activations and the impact of gravity on tongue positioning”. In: *J Acoust Soc Am* 126 (2009).

[11] S. S. Blemker et al. “A 3D model of muscle reveals the causes of nonuniform strains in the biceps brachii”. In: *J Biomech* 38 (2005), pp. 657–665.

[12] I. Stavness et al. “Coupled hard-soft tissue simulation with contact and constraints applied to jaw-tongue-hyoid dynamicsl”. In: *Int J Numer Method Biomed Eng* 27 (2011), pp. 367–390.

[13] A. Top et al. “Active learning for interactive 3D image segmentation”. In: *Proc Med Image Comput Comput Assist Interv (MICCAI)*. 2011, pp. 603–610.

[14] M. Bucki et al. “A fast and robust patient specific finite element mesh registration technique: application to 60 clinical cases”. In: *Med Image Anal* 13 (2010), pp. 303–317.

[15] A. Myronenko and X. Song. “‘Point set registration: Coherent point drift”. In: *IEEE Trans Pattern Anal Mach Intell* 32 (2010), pp. 2262–2275.

[16] I. Stavness et al. “Unified skinning of rigid and deformable models for anatomical simulations”. In: *Tech Brief SIGGRAPH Asia*. 2014, p. 9.