Physics-Based Deformable Tongue Visualization

Yin Yang, *Student Member*, *IEEE*, Xiaohu Guo, *Member*, *IEEE*, Jennell Vick, Luis G. Torres, *Student Member*, *IEEE*, and Thomas F. Campbell

Abstract—In this paper, a physics-based framework is presented to visualize the human tongue deformation. The tongue is modeled with the Finite Element Method (FEM) and driven by the motion capture data gathered during speech production. Several novel deformation visualization techniques are presented for in-depth data analysis and exploration. To reveal the hidden semantic information of the tongue deformation, we present a novel physics-based volume segmentation algorithm. This is accomplished by decomposing the tongue model into segments based on its deformation pattern with the computation of *deformation subspaces* and fitting the target deformation locally at each segment. In addition, the strain energy is utilized to provide an intuitive low-dimensional visualization for the high-dimensional sequential motion. Energy-interpolation-based morphing is also equipped to effectively highlight the subtle differences of the 3D deformed shapes without any visual occlusion. Our experimental results and analysis demonstrate the effectiveness of this framework. The proposed methods, though originally designed for the exploration of the tongue deformation, are also valid for general deformation analysis of other shapes.

Index Terms—Deformable model, tongue, finite element method, modal analysis

1 Introduction

A realistic and effective visualization of the human tongue for specific speech tasks is of importance in the domain of speech research and has numerous potential applications. For example in the rehabilitation of speech disorders, the visualized 3D tongue motion could provide a visual model that would help an individual achieve the correct position of the tongue during the production of various speech sounds.

The tongue is an interior muscular organ and its movement during the production of speech sounds is subtle and swift. A complete production of an individual Consonant-Vowel (CV) syllable takes only tenths of a second. Hence, ordinary optical devices or imaging modalities such as video cameras or CT imaging may not be sufficient for the real-time motion acquisition of the tongue. Furthermore, creating a realistic visual representation of tongue movement based on the collected experimental data is also a challenging problem. The anatomical structure of the tongue is quite complex, requiring four intrinsic and four extrinsic muscles to be tightly coordinated during normal speech sound production [29], [40]. The question of what specific mechanism is responsible for the precise coordination of tongue movement during speech production continues to be of significant interest for speech researchers. As a result, modeling tongue motion for a specified speech task is not common in previous research.

To address these challenges, we propose a novel visualization framework which combines the motion capture (*mocap* for short in the rest of this paper) technique and the physics-based modeling strategy. The movement of the tongue was tracked with mocap sensors with high sampling frequency (100 Hz). The number of available mocap sensors that we can place on the subject was limited as too many sensors placed on the tongue could severely impede the regular pronunciation. Thus, the collected mocap data are not able to represent the entire 3D geometry of the tongue. Fortunately, this disadvantage is compensated for by the adopted physics-based deformable model. The deformation of the tongue can be simulated with the appropriate external stimuli, the activations from various muscle groups [35], [41]. However, detecting and recording internal muscle tension would require invasive instrumentation, (e.g., hook-wire EMG) that is not feasible for most applications. Alternatively, we embed the collected mocap data into the finite element mesh of the tongue and deform the mesh so that the nodes corresponding to the sensors comply with the real mocap data. The necessary constraint forces applied to the constrained nodes (which correspond to the unknown *Lagrange multipliers* as in [34]), instead of muscular activations, drive forward the deformation.

We also present a novel deformable shape segmentation algorithm in this framework to reveal the in-depth semantic information of the tongue deformation. This technique segments a given deformable model into partitions based on its deformation pattern. The segmentation is achieved by constructing *deformation subspaces* and fitting the segmented deformable model to the target deformation locally within the subspaces. The users can comprehend the "trend" of the deformation and some in-depth analysis is made possible

- *Y. Yang and X. Guo are with the Department of Computer Science, The University of Texas at Dallas, 800 West Campbell Road, Richardson, TX 75080-3021. E-mail: {yinyang, xguo}@utdallas.edu.*
- *J. Vick is with the Department of Psychological Sciences, Case Western Reserve University. E-mail: jennell@case.edu.*
- *L.G. Torres is with the Department of Computer Science, The University of North Carolina at Chapel Hill. E-mail: luis@cs.unc.edu.*
- *T.F. Campbell is with Callier Center for Communication Disorders, The University of Texas at Dallas, Richardson, TX 75080-3021. E-mail: tfc061000@utdallas.edu.*

Manuscript received 13 Aug. 2011; revised 27 Feb. 2012; accepted 14 May 2012; published online 22 Aug. 2012.
Recommended for acceptance by H.-S. Ko.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org, and reference IEEECS Log Number TVCG-2011-08-0185.
Digital Object Identifier no. 10.1109/TVCG.2012.174.

based on the visualized information. For example, communication scientists could identify the regional behavior abnormality of the tongue more easily with this technique.

The elastic energy of the deformation is leveraged as a pseudoshape metric mapping the high-dimensional mesh motions to the 2D energy-time domain for a low-dimensional visual representation. With this energy-based visualization, users can easily obtain a general understanding of the differences among various tongue deformations. The more detailed 3D shape comparisons between two specific frames are visualized with a novel morphing algorithm. The two 3D deformed shapes under comparison, namely, the *source deformation* and the *target deformation* constitute the first and last frame of the animation and the deformation morphs from the source to the target smoothly. This guarantees an occlusion-free visual perception of the shape differences. Furthermore, a layout with fully synchronized 2D, 3D, as well as side-by-side multiple small views is adopted which delivers comprehensive visual feedback to the user.

Overall, the contributions of this work can be summarized as follows:

- A physics-based deformable tongue visualization framework is presented using the finite element method (FEM) driven by mocap data. This approach constructs an intuitive 3D animation of the tongue for a specified speech task and delivers enriched visual information.
- A novel deformable shape segmentation algorithm is presented in this framework. The segmentation is computed by constructing deformation subspaces at each segment and fitting the segmented model to the target deformation locally within the subspaces. We extend the *local rigid subspace* to the *local constraint subspace*, which ensures the interface compatibility between segments. It can be further extended to incorporate additional user-specified constraints of the deformable model.
- A comparative visualization of deformation sequences based on the elastic energy is also presented. In addition, a deformation morphing tool is integrated into the proposed framework, which can generate a smooth animation to enable pairwise visual comparison between 3D deformations with an occlusion-free fashion.

Although we are targeting tongue deformation analysis in this paper, the proposed visualization techniques can also be used for other applications involving general deformable shape analysis.

The rest part of the paper is organized as follows. Section 2 briefly reviews the related literature. Section 3 describes the gathering and preprocessing of the mocap data. Section 4 summarizes the physics-based deformation modeling algorithm adopted in this framework. Section 5 presents a new deformation segmentation method in detail which is helpful for domain experts to intuitively identify general deformation pattern of the deformed tongue shape. In Section 6, experimental results and analysis are highlighted with the introduction of some additional features of the framework, (e.g., the low-dimensional energy-based visualization and morphing). Finally, some potential future works are discussed in Section 7.

2 Related Work

The human tongue is critical for speech production. Investigating its behavior and contribution to speech production has been of interest to researchers in linguistics, phonetics, and physiology. Magnetic Resonance Imaging (MRI) has been used as the data source [2], [3] for reconstructing static 3D tongue shapes. To capture the tongue motion, three sagittal directions of MRI images [37] have been used to record the 2D contours of tongue in three sagittal planes. However, the MRI acquisition frequency is still too low for catching the subtle tongue motion during real speech production. In addition, it is also very difficult to reconstruct the complete 3D shapes of the tongue, since its shapes are only imaged at three sagittal planes. X-ray imaging systems [36] have higher temporal resolution. However, they expose the speaker to radiation. Ultrasound systems [33], [36] have also been widely used for modeling tongue movements. However, they are unable to track the motion of the tongue tip [22], [38] because of the air gaps that surround it.

Biomechanical models of the tongue with physics-based FEM is a widely used method [5], [13], [14], [42]. In most existing literature [10], [13], [35], [41], the deformation of the tongue is driven by muscle activations. Although interesting results have been reported with this method, it is still difficult to simulate the tongue motion for the specified speech tasks, since it is very challenging to figure out how the different muscle groups act for various speech productions. Conversely, inverse-dynamics technique that builds the unknown motion based on preknown constraints [34] has obvious advantages on this challenge. With this technique, the mocap data can be used as position constraints to the deformable model: the nodes on the tongue mesh corresponding to the mocap sensors always mirror the sensors' movement, while continuum mechanics governs the deformation of the rest of the tongue to create a convincing visual simulation of the tongue's movement. In fact, the constraint-driven deformable model has been widely used for computer animation [24], [43].

Recent research indicates that the deformation of the tongue could be as large as 200 percent in compression and 160 percent in elongation [12], which suggests the linear finite element method based on the hypothesis of small deformation may not be appropriate. The linear strain tensor is commonly used in deformable modeling for its computational efficiency. This type of deformable model is known as *linear elasticity* or *linear deformation*. A well-known drawback associated with the usage of the linear tensor is that it is inept for rotational deformation as the linear strain tensor generates inappropriate strains when the object undergoes rotation. *Stiffness warping* [25] is a good solution to this problem, where a local nodal rotation component is evaluated and added back to the dynamical motion integral. Existing literature of tongue modeling shows that the tongue deformation is locally smooth and continuous during speech. This finding indicates the potential feasibility of using modal reduction [28] to speed up the simulation, since the high-frequency vibrations are not likely to occur in speech production. A real-time computed tongue deformation could make the framework more practical for real applications such as language training. *Modal warping* [7] unites the benefits of both modal reduction and stiffness warping. The *curl* of the deformation field is calculated and used to warp the linear

Fig. 1. The placement of the mocap sensors on the subject during data collection.

deformation. This warping technique excels in its independence to time integration and provides an expedient map from any linear deformation to a well-estimated warped deformation. This technique is adopted in our framework to simulate the deformation of the tongue.

Another innovative feature of our framework is the equipped tool of deformation segmentation. Although many works have been published on mesh segmentation [31], most existing works are based on static geometry by computing certain feature descriptors targeting the surface mesh. Our algorithm differs from these works as it is applied to the physics-based deformation sequence and segments the volumetric mesh according to local deformation subspace. Local deformation modes [8] are adopted to construct the low-ranked subspace of deformation. The subspace spans deformations with less variation, which are considered "uniform." Huang et al. [19] used only the low-frequency eigenmodes of the Hessian of a deformation energy to span the deformation subspace, which is fundamentally equivalent to the use of the rigid subspace in this paper. We extend the rigid subspace to constraint subspace, so that the segment interface compatibility and users' selected constraints can be accommodated. The computed segmentation highlights local semantic information behind the tongue deformation sequences.

Occlusion-free visual comparison between deformations is also made possible with deformation morphing. As deformation is associated with 3D geometry, occlusion is almost inevitable when multiple deformations are displayed together. Even with multiple side-by-side viewports rendering individual deformations simultaneously, subtle differences in deformations may still be difficult to visually discern. Our method complements other physics-based morphing algorithms [4], [17], [18], [44] as it is based on modal displacement interpolation and integrated into modal reduction with high computational efficiency. For shape morphing, rotation typically needs to be separately handled [4], [39]. Similarly, we only blend linear modal deformations and the rotation is computed based on the blended linear part afterward.

3 Motion Data Collection and Preprocessing

Based on the literature review [35], [37], [38] we believe that the higher strain is likely to happen at the lateral edges of the tongue blade and the midsagittal direction. Accordingly, mocap sensors were placed on the surface of the tongue in a rhombus configuration as shown in Fig. 1. Specifically, two sensors were placed at the midline $1\,\mathrm{cm}$ and $3\,\mathrm{cm}$ from the tip of the tongue. Two other sensors

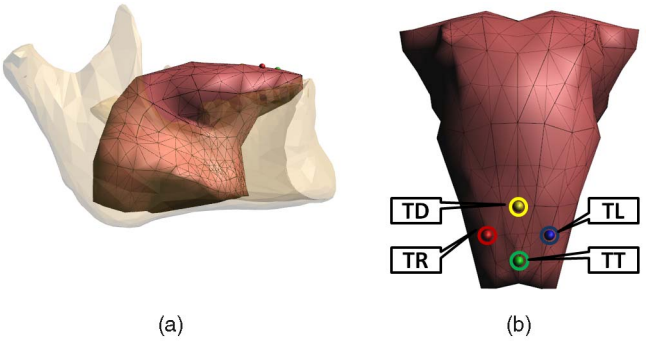


Fig. 2. (a) The tetrahedral mesh used for the simulation and its spatial relationship to the jaw shown as transparent. (b) The nodes corresponding to the mocap sensors.

were placed $2\,\mathrm{cm}$ from the tip of the tongue and $1\,\mathrm{cm}$ from the left and right edge of the tongue, respectively. In this study, we only focus on the tongue's relative deformation to the jaw. Therefore, the motion of the jaw was subtracted from the mocap data. As reported in literature [27], the jaw's motion leads to a rigid translation. Accordingly, this translation was tracked with an additional sensor placed at the mandibular symphysis. The mocap system used in our data collection was *Wave Speech Research System* from *Northern Digital, Inc.* [26]. The motion of the sensors were tracked at $100\,\mathrm{Hz}$. Participants read instructions for the speech tasks from a screen in their visual field. Elicited tasks were CV syllable trains that were five syllables in length, (e.g., /ta-ta-ta-ta-ta/) and each train was repeated five times. Consonants produced includes /t/, /l/, /k/, /r/. The vowel in each syllable train was /a/. Audio was recorded using a head-mounted microphone and the acoustic signals were temporally synchronized with the x, y, and z coordinates of the tracked sensor motion. There were totally seven subjects who participated in the data collection including one male and one female native English speakers and four male speakers whose first language is not English. In addition, the tongue motion data were also collected from one adult female with a speech disorder (dysfluency).

3.1 Spatial Data Alignment

Instead of adapting the finite element mesh to real speakers' tongue geometry [13], the raw mocap data were scaled to a "standard" tetrahedral mesh. Simulating deformation on one single target mesh facilitates shape comparison and analysis. However, it could also induce simulation errors due to the geometrical inconsistence between the used mesh and the real tongue of the subjects. The finite element mesh used in our framework is from *ArtiSynth* [9] which consists of 1,803 nodes and 8,606 tetrahedral elements (Fig. 2a). Four sensors, $\mathbf{S}_{TT}$, $\mathbf{S}_{TD}$, $\mathbf{S}_{TL}$, $\mathbf{S}_{TR}$ are associated with four nodes, $\mathbf{P}_{TT}$, $\mathbf{P}_{TD}$, $\mathbf{P}_{TL}$, and $\mathbf{P}_{TR}$ on the tetrahedral mesh (Fig. 2b). Each participant was asked to leave the tongue at rest for a short period of time (about 5 seconds) and the average positions of the sensors are used as the reference configuration of the tongue denoted with the superscript $rest$. At each frame of the mocap data, the contribution of the jaw motion is removed by subtracting the jaw's displacement

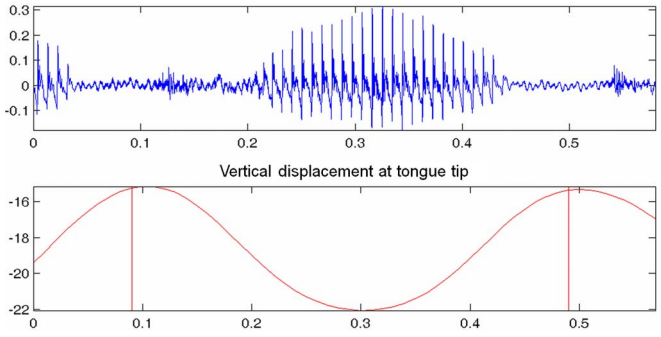


Fig. 3. Temporal segmentation of the mocap data corresponding to /ta/. Top: audio waveform. Bottom: the vertical displacement of the tongue tip sensor. Vertical lines are zero crossings in the velocity record to assist with parsing.

with respect to its rest position at sensor $\mathbf{S}_J$, the fifth sensor placed at the mandibular symphysis.

$$\mathbf{S}^i \leftarrow \mathbf{S}^i - (\mathbf{S}^i_J - \mathbf{S}^{rest}_J),$$

where superscript i simply denotes the frame index. After that, a translation is applied so that the center of the four sensors at rest positions and the center of their corresponding nodes on the mesh overlap.

Next, two rotations pivoted at the center of the sensors are applied to align $\mathbf{S}^{rest}_{TT} - \mathbf{S}^{rest}_{TD}$ with $\mathbf{P}_{TT} - \mathbf{P}_{TD}$ as well as $\mathbf{S}^{rest}_{TL} - \mathbf{S}^{rest}_{TR}$ with $\mathbf{P}_{TL} - \mathbf{P}_{TR}$. Finally, a uniform scaling, s is performed so that the sum of the distance between each sensor and its corresponding node on the tetrahedral mesh is minimized

$$\arg\min_s \sum \|s\mathbf{S}^{rest} - \mathbf{P}\|,$$

which can be easily computed by taking the first order derivative of the target function with respect to s. A homogeneous matrix can be assembled composing all the mentioned transformations which converts the raw mocap data into the coordinate system of the tetrahedral mesh.

3.2 Temporal Segmentation

The raw mocap data were segmented into pieces such that each piece corresponds to a complete production of one CV. With the help of the recorded acoustic signal, each syllable train was first identified and roughly parsed. Individual productions of the CVs were more finely segmented using the vertical displacement and its corresponding velocity record from the most anterior tongue sensor ($\mathbf{S}_{TT}$). Consonant release is defined as the velocity zero crossing corresponding to the local maxima in the vertical displacement record and the vowel displacement is defined as the velocity zero crossing corresponding to the local minima [15]. A full production was segmented from one consonant release to another as shown in Fig. 3, where two red vertical lines indicate the first and last frame for one complete motion sequence corresponding to one production of /ta/. The corresponding acoustic signals are shown in the top of the figure. Only the second, third, and fourth out of the five repeated syllables in each train were used because the first and last productions in a train were produced with larger displacements and longer durations. The processed mocap data are used to drive the finite element mesh simulated with modal warping.

4 Deformation Modeling

In this section, we briefly review the linear deformable model with modal analysis and modal warping while the detailed derivations can be found in the related literature, (e.g., [7], [28]).

The motion of the deformable tongue discretized with finite element mesh can be expressed with *Euler Lagrange equation* [11]:

$$\mathbf{M}\ddot{\mathbf{u}} + \mathbf{C}\dot{\mathbf{u}} + \mathbf{K}\mathbf{u} = \mathbf{f}, \tag{1}$$

where $\mathbf{M}$, $\mathbf{C}$, and $\mathbf{K}$ are the mass, damping, and stiffness matrices determined by the material's intrinsic physical properties. Previous research [12], [13], [14], [41], indicates that an isotropic modeling for the tongue is applicable as the variation of Young's modulus at different parts of the tongue is very small [13]. The material parameters in our framework follow the experimental work reported in [12] and [41]. Specifically, the Young's modulus is set as 6,912 and the Poisson's ratio is set as 0.49. The column vectors $\mathbf{u}$ and $\mathbf{f}$ represent concatenated nodal displacements and external forces. If we solve a generalized eigenproblem, $\mathbf{K}\phi = \mathbf{M}\phi\lambda$, and assemble a matrix $\mathbf{\Phi}$ with the m eigenvectors corresponding to first m smallest eigenvalues, a low-frequency free vibration space (or called *modal space*) of rank m can be constructed. The *modal displacement*, $\mathbf{q}$ is related with its spatial counterpart by $\mathbf{u} = \mathbf{\Phi}\mathbf{q}$. Substituting it into (1) followed by a premultiplication of $\mathbf{\Phi}^\top$ leads to

$$\mathbf{M}_q\ddot{\mathbf{q}} + \mathbf{C}_q\dot{\mathbf{q}} + \mathbf{K}_q\mathbf{q} = \mathbf{f}_q, \tag{2}$$

which can be considered as a modal version of (1). $\mathbf{M}_q$ is an identity matrix and $\mathbf{K}_q$ is a diagonal matrix of eigenvalues. $\mathbf{C}_q$ can be considered as a linear combination of $\mathbf{M}_q$ and $\mathbf{K}_q$ under *Rayleigh damping* condition, e.g., $\mathbf{C}_q = \alpha\mathbf{M}_q + \beta\mathbf{K}_q$. α and β are set as 6.22 and 0.11, respectively, as reported in [12] and [41]. In order to accommodate rotational deformation with linear strain tensor, Choi and Ko [7] proposed a modal version of corotational deformable model called *modal warping*. At each time frame t, the nonlinear deformation $\tilde{\mathbf{u}}(t)$ is approximated as the accumulation of a linear deformation mended by a small local rotation: $\tilde{\mathbf{u}} = \tilde{\mathbf{R}}\mathbf{\Phi}\mathbf{q}$, where $\tilde{\mathbf{R}}$ is the accumulated nodal rotation.

Decoupling (2) results in a linear system with the form of $\mathbf{A}\ddot{\mathbf{q}} = \mathbf{b}$. For instance, the *Newmark-average acceleration method* [20] will have $\mathbf{A} = \mathbf{M}_q + \frac{h}{2}\mathbf{C}_q + \frac{h^2}{4}\mathbf{K}_q$ and $\mathbf{b} = \mathbf{f}_q - \mathbf{C}_q\dot{\mathbf{q}}^* - \mathbf{K}_q\mathbf{q}^*$, where $\dot{\mathbf{q}}^* = \dot{\mathbf{q}}^- + \frac{h}{2}\ddot{\mathbf{q}}^-$ and $\mathbf{q}^* = \mathbf{q}^- + h\dot{\mathbf{q}}^- + \frac{h^2}{4}\ddot{\mathbf{q}}^-$ are two *predictors*. The superscript "$-$" denotes the values in previous time frame. The modal displacements and velocities can be computed through

$$\begin{cases} \mathbf{q} = \mathbf{q}^* + \dfrac{h^2}{4}\ddot{\mathbf{q}}, \\ \dot{\mathbf{q}} = \dot{\mathbf{q}}^* + \dfrac{h}{2}\ddot{\mathbf{q}}, \end{cases} \tag{3}$$

h is the time interval between two frames and set as 0.01 second according to the actual data sampling rate of the mocap sensor.

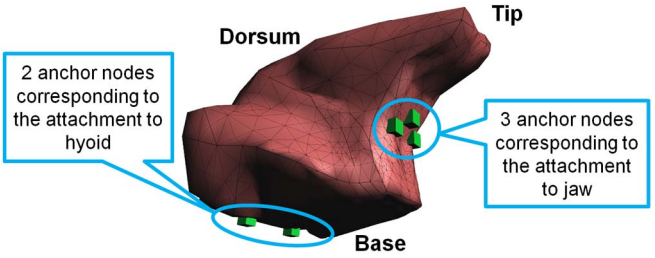


Fig. 4. Five anchor nodes on the tongue mesh at the rest configuration.

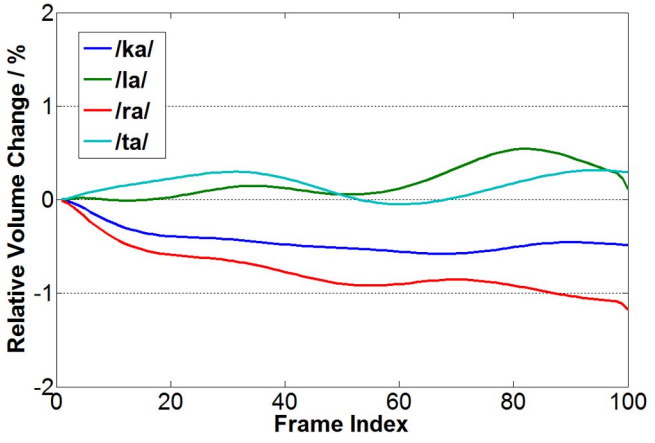


Fig. 5. Average volume change for each CV syllable.

Nodes corresponding to sensors are constrained according to the normalized mocap data such as $\tilde{\mathbf{R}}_s\boldsymbol{\Phi}_s\mathbf{q} = \mathbf{c}$, where $\tilde{\mathbf{R}}_s$ and $\boldsymbol{\Phi}_s$ are submatrices of $\tilde{\mathbf{R}}$ and $\boldsymbol{\Phi}$ corresponding to the Degrees Of Freedom (DOFs) of sensor nodes and $\mathbf{c}$ is the values which the constrained DOFs must satisfy with. In addition to the sensor nodes, we also set five anchor nodes on the mesh as in Fig. 4. The anchor nodes are fixed during the simulation. They are chosen corresponding to two major attachments of the tongue: the hyoid and jaw. All the linear constraint equations can be integrated into (2) with *Lagrange multiplier method*.

In reality, the volume of the tongue tissue remains constant during speech as it is composed mostly of water, which is incompressible. As a result, the incompressible constraint is sometimes adopted in previous published works [35]. However, this nonlinear constraint is not applied in our simulation as our experiments indicate that the volume change with modal warping is very small (less than 1 percent in most of the cases as shown in Fig. 5).

5 Segmentation of Deformed Shapes

For a simple 2D example of a curve as shown in Fig. 6, one may fast identify that the target curve roughly has three parts and each part can be well approximated by a straight line segment. When targeting a complex 3D deformation over volumetric mesh consisting of thousands of elements, demonstrating such regional deformation feature is obviously much more challenging. However, domain scientists often expect such information. For instance, they may suspect the subject has a certain unique tongue movement due to the native language, gender or pathological abnormality when producing a particular speech sound. Effectively highlighting and visualizing such information could be of help for the in-depth speech analysis. Motivated

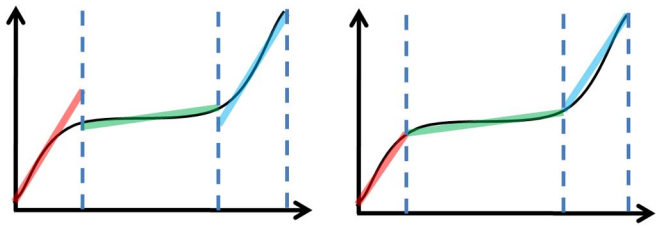


Fig. 6. Simple segmentation examples on a 2D curve.

by such requirement, we propose a novel physics-based deformation segmentation algorithm. Similar to using the line segments to fit the curve as in Fig. 6, we use the low-ranked deformation to fit the target deformation locally and find out the segmentation with the best approximation.

Before the detailed description of the segmentation algorithm, several terminologies are defined first for the further convenience: 1) a *partition* $\mathcal{P}$ is a face-connected subset of a tetrahedral mesh $\mathcal{T}$. The face-connection between elements means two tetrahedra are connected if and only if they are sharing one triangle face; 2) a *one-ring-neighbor* of $\mathcal{P}$, denoted by $\mathcal{N}_1^{\mathcal{P}}$, is a set of tetrahedra that do not belong to any partitions such that any tetrahedron in $\mathcal{N}_1^{\mathcal{P}}$ is connected with a tetrahedron in $\mathcal{P}$.

The general segmentation, as illustrated in Algorithm 5.1, begins with several initial seeding partitions $\mathcal{S}_i$ specified by the user. All the partitions are expanded by including their one-ring-neighbors. An approximated deformation $\tilde{\mathbf{u}}_{\mathcal{P}}$ is computed to locally fit the target deformation $\mathbf{u}_{\mathcal{P}}$ at each expanded partition. $\tilde{\mathbf{u}}_{\mathcal{P}}$ is computed within a low-ranked subspace which suggests that the DOFs associated with $\tilde{\mathbf{u}}_{\mathcal{P}}$ is small. The target deformation is considered locally uniform if it is well fitted by $\tilde{\mathbf{u}}_{\mathcal{P}}$. The quality of the local fitting is measured with the partition approximation error $e_{\mathcal{P}}$ that can be computed as

$$e_{\mathcal{P}} = \frac{\|\tilde{\mathbf{u}}_{\mathcal{P}} - \mathbf{u}_{\mathcal{P}}\|^2}{V_{\mathcal{P}}}, \tag{4}$$

where $V_{\mathcal{P}}$ is the volume of the partition at the rest shape. Only the expansion of the partition with minimum error is kept while all the rest partitions rollback to their unexpanded status. The competition for expansion among partitions continues till every element of the mesh is assigned with a partition.

Algorithm 5.1. Greedy Segmentation

for all $\mathcal{P}_i$ **do**
  $\mathcal{P}_i \leftarrow \mathcal{S}_i$ // *initialization*
end for
while $\mathcal{T}$ has un-partitioned tetrahedra **do**
  for all $\mathcal{P}_i$ **do**
    $\mathcal{P}_i \leftarrow \mathcal{P}_i + \mathcal{N}_1^{\mathcal{P}_i}$
    $\mathcal{M}_i \leftarrow \mathcal{N}_1^{\mathcal{P}_i}$ // *backup* $\mathcal{N}_1^{\mathcal{P}_i}$
    compute $\tilde{\mathbf{u}}_{\mathcal{P}_i}$ with local deformation modes
    compute $e_{\mathcal{P}_i}$
  end for
  $j \leftarrow i|min(e_{\mathcal{P}_i})$
  for all $i, i \neq j$ **do**
    $\mathcal{P}_i \leftarrow \mathcal{P}_i - \mathcal{M}_i$ //*rollback partitions*
  end for
end while

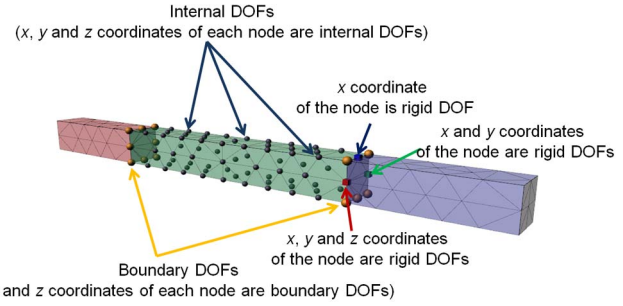


Fig. 7. DOF classification of the green partition of the bar model. All the x, y, and z freedoms of the purple nodes are internal DOFs and all the freedoms of the yellow nodes are boundary DOFs. The rest three nodes represented with red, green, and blue cubes are the nodes containing rigid DOFs.

We can also apply the algorithm on a sequence of deformation, (e.g., a complete tongue motion corresponding to one CV production). In this case, $\tilde{\mathbf{u}}_{\mathcal{P}}$ is computed for each frame of the motion and the error evaluation (4) becomes an energy-weighted summation over the whole sequence

$$e_{\mathcal{P}} = \frac{\sum_i (U^i \|\tilde{\mathbf{u}}^i_{\mathcal{P}} - \mathbf{u}^i_{\mathcal{P}}\|^2)}{V_{\mathcal{P}} \sum_i U^i}, \tag{5}$$

where the superscript i indicates the frame index. U is the elastic energy of the deformed mesh which can be computed as $U = \frac{1}{2}\mathbf{u}^\top \mathbf{K}\mathbf{u}$. Under modal analysis, the computation of U is significantly simplified if we replace $\mathbf{u}$ with $\mathbf{\Phi q}$ which yields

$$U = \frac{1}{2}\mathbf{q}^\top \mathbf{K}_q \mathbf{q}. \tag{6}$$

In order to compute $\tilde{\mathbf{u}}_{\mathcal{P}}$, a low-rank local deformation subspace needs to be constructed. The basis vectors of the subspaces are called (local) *modes* which are essentially the pre-computable displacements of the partitions. In the following sections, we present two different types of local modes, the *rigid mode* and the *constraint mode*, respectively.

5.1 Segmentation with Rigid Modes

The term rigid motion/displacement refers to a class of displacements that do not induce any strain over the mesh. All the rigid displacements of a partition $\mathcal{P}$ can be represented using its rigid modes $\mathbf{\Phi}^{\mathcal{P}}_{rig}$.

The rigid modes are the partition's displacements corresponding to the system response to the external stimulus on the *statically determinate* or rigid DOFs of $\mathcal{P}$. We denote the set including all DOFs of the partition as $\mathbb{F}$ and the size of $\mathbb{F}$ is $n_{\mathbb{F}}$. For a partition with k nodes, $n_{\mathbb{F}}$ is $3k$ as each node has x, y, and z independent freedoms. Six statically determinate DOFs (denoted by $\mathbb{R}$) can be picked out with the following steps: for a boundary triangle face on $\mathcal{P}$ with nodes $\mathbf{n}_0, \mathbf{n}_1$, and $\mathbf{n}_2$, the first three DOFs are the x, y, z coordinates of $\mathbf{n}_0$; the 4th and 5th DOFs are chosen as x- and y- coordinates of $\mathbf{n}_1$; and the last DOF is either x- or y-coordinate of $\mathbf{n}_2$ (Fig. 7). If we slowly[1] apply a unit displacement to a DOF in $\mathbb{R}$ while keeping the other five DOFs in $\mathbb{R}$ fixed, a set of static equilibrium equations can be listed correspondingly (in order to simplify notation, the superscript $\mathcal{P}$ indicating local variable for partition $\mathcal{P}$ is omitted in the following formulations):

$$\begin{bmatrix} \mathbf{K}_{\bar{\mathbb{R}}\bar{\mathbb{R}}} & \mathbf{K}_{\bar{\mathbb{R}}\mathbb{R}} \\ \mathbf{K}_{\mathbb{R}\bar{\mathbb{R}}} & \mathbf{K}_{\mathbb{R}\mathbb{R}} \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_{\bar{\mathbb{R}}} \\ \mathbf{I}_{\mathbb{R}} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \tag{7}$$

where $\bar{\mathbb{R}}$ is the complement set of $\mathbb{R}$ such that $\bar{\mathbb{R}} = \mathbb{F} - \mathbb{R}$ and the stiffness matrix of the partition $\mathbf{K}$ is reorganized according to the classification of the DOFs as the indexing order of nodes on a finite element mesh is arbitrary. $\mathbf{I}_{\mathbb{R}}$ is a 6×6 identity matrix corresponding to the imposed unit displacements on $\mathbb{R}$. It is noteworthy that the rigid mode itself is a zero-strain displacement over the partition. Thus, (7) has the form of $\mathbf{K\Phi}_{rig} = \mathbf{0}$ without any external force on the right-hand side of the equation. $\mathbf{\Phi}_{\bar{\mathbb{R}}}$ is the unknown responsive displacements at $\bar{\mathbb{R}}$ and can be computed by expanding only the first line of (7). Therefore, $\mathbf{\Phi}_{rig}$ can be computed through

$$\mathbf{\Phi}_{rig} = \begin{bmatrix} \mathbf{\Phi}_{\bar{\mathbb{R}}} \\ \mathbf{I}_{\mathbb{R}} \end{bmatrix} = \begin{bmatrix} -\mathbf{K}^{-1}_{\bar{\mathbb{R}}\bar{\mathbb{R}}} \mathbf{K}_{\bar{\mathbb{R}}\mathbb{R}} \\ \mathbf{I}_{\mathbb{R}} \end{bmatrix}. \tag{8}$$

For a given target deformation $\mathbf{u}$, the approximated rigid displacement $\tilde{\mathbf{u}}_{rig}$ that minimizes $e_{\mathcal{P}}$, or equivalently $\min\|\mathbf{u} - \tilde{\mathbf{u}}_{rig}\|^2$, can be computed with

$$\tilde{\mathbf{u}}_{rig} = \mathbf{\Phi}_{rig}(\mathbf{\Phi}^\top_{rig}\mathbf{\Phi}_{rig})^{-1}\mathbf{\Phi}^\top_{rig}\mathbf{u}. \tag{9}$$

1. "Slow" indicates the system response is static without the consideration of inertia or acceleration.

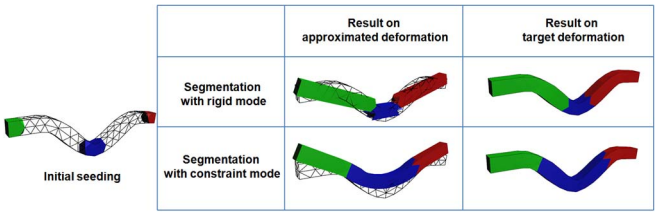


Fig. 8. Segmentation results with rigid modes and constraint modes on the bar whose middle portion is bent.

The first row in Fig. 8 shows the segmentation results on an example bar model. The target deformation is a bending deformation at the middle part of the bar. Fitted with local rigid modes, the target deformation is segmented into three partitions based on user-specified initial seedings. At each partition, a rigid body motion is used to approximate the target deformation.

5.2 Segmentation with Constraint Modes

One obvious defect of rigid segmentation as we can see from Fig. 8 is that the approximated rigid displacement do not satisfy the interface compatibility. If we look back at our 2D curve example as mentioned at the beginning of this section, this is analogous to fitting the target curve using separate line segments (Fig. 6left). In some circumstances, a continuous approximation is preferred, (e.g., Fig. 6right). Hence, strain-incorporated local modes should be used instead of the rigid modes.

As shown in Fig. 7, let $\mathbb{B}$ denote the partition's *boundary* DOFs and its complement set is symbolled with $\mathbb{I}$ holding only *internal* DOFs. The sizes of $\mathbb{I}$ and $\mathbb{B}$ are $n_{\mathbb{I}}$ and $n_{\mathbb{B}}$, respectively, and $n_{\mathbb{I}} + n_{\mathbb{B}} = n_{\mathbb{F}}$. Similar to rigid modes, we slowly apply a unit displacement to one DOF in $\mathbb{B}$ while fixing the rest DOFs in $\mathbb{B}$. The constraint modes $\mathbf{\Phi}_{con}$ are the

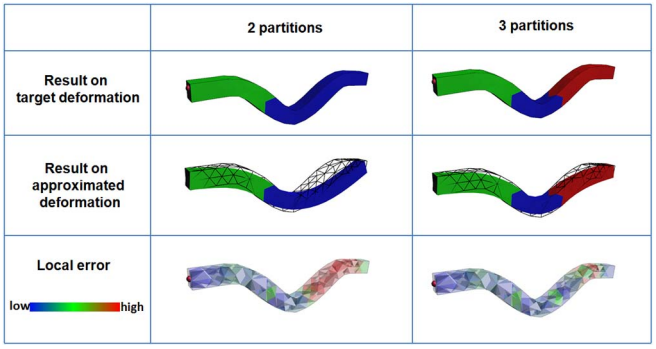


Fig. 9. Segmentation results using autoseeding with user-specified constraints. Two constrained nodes are added at both ends of the bar model. New seeding is chosen based on the local error after prior segmentation. The approximated deformation always coincides with the target deformation at the constrained nodes.

corresponding responsive displacements of the partition which can be computed by solving the following equilibrium equations:

$$\begin{bmatrix} \mathbf{K}_{\mathbb{II}} & \mathbf{K}_{\mathbb{IB}} \\ \mathbf{K}_{\mathbb{IB}\mathbb{I}} & \mathbf{K}_{\mathbb{IB}\mathbb{IB}} \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_{\mathbb{I}} \\ \mathbf{I}_{\mathbb{IB}} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{f}_{\mathbb{IB}} \end{bmatrix}, \tag{10}$$

which yields

$$\mathbf{\Phi}_{con} = \begin{bmatrix} \mathbf{\Phi}_{\mathbb{I}} \\ \mathbf{I}_{\mathbb{IB}} \end{bmatrix} = \begin{bmatrix} -\mathbf{K}_{\mathbb{II}}^{-1}\mathbf{K}_{\mathbb{IIB}} \\ \mathbf{I}_{\mathbb{IB}} \end{bmatrix}. \tag{11}$$

where $\mathbf{I}_{\mathbb{IB}}$ is a $n_{\mathbb{IB}} \times n_{\mathbb{IB}}$ identity matrix. An approximated deformation in constraint space $\tilde{\mathbf{u}}_{con}$ can be expressed with constraint modes

$$\tilde{\mathbf{u}}_{con} = \begin{bmatrix} \tilde{\mathbf{u}}_{\mathbb{I}} \\ \tilde{\mathbf{u}}_{\mathbb{IB}} \end{bmatrix} = \mathbf{\Phi}_{con}\mathbf{q}_{con} = \begin{bmatrix} \mathbf{\Phi}_{\mathbb{I}} \\ \mathbf{I}_{\mathbb{IB}} \end{bmatrix} \mathbf{q}_{con}, \tag{12}$$

where $\mathbf{q}_{con}$ is the generalized *constraint coordinate*. For a given target deformation $\mathbf{u}$, it is expected that the approximated deformation overlaps the target deformation at the boundary DOFs or, equivalently $\tilde{\mathbf{u}}_{\mathbb{IB}} = \mathbf{u}_{\mathbb{IB}}$. Expanding the bottom row of (12) yields $\mathbf{q}_{con} = \mathbf{u}_{\mathbb{IB}}$. Hence, the approximated constraint deformation can be computed through

$$\tilde{\mathbf{u}}_{con} = \mathbf{\Phi}_{con}\mathbf{u}_{\mathbb{IB}}. \tag{13}$$

The formulation of rigid modes (7) and constraint modes (10) are very similar except that the triggered DOFs are extended from statically determinate DOFs $\mathbb{IR}$ to the entire boundary DOFs $\mathbb{IB}$. The strain is also generated in constraint equilibrium. Accordingly external forces at the boundary $\mathbf{f}_{\mathbb{IB}}$ are necessary to move or fix the boundary DOFs as in the right-hand-side of (10).

Additional user-specified constraints can also be incorporated into the segmentation. An example is the anchor nodes on the tongue (Fig. 4): it is desirable that the approximated deformations also have zero displacements at these anchor nodes. Denote $\mathbb{UW}$ as the set of freedoms that are constrained by the user. We can simply merge $\mathbb{UW}$ into $\mathbb{IB}$ such that $\mathbb{IB} \leftarrow \mathbb{IB} + \mathbb{UW}$ and $\mathbb{II} \leftarrow \mathbb{II} - \mathbb{UW}$. Because the boundary freedoms always coincide with the target deformation, the user-specified constraints will automatically be satisfied when $\mathbb{UI}$ is classified as boundary DOFs.

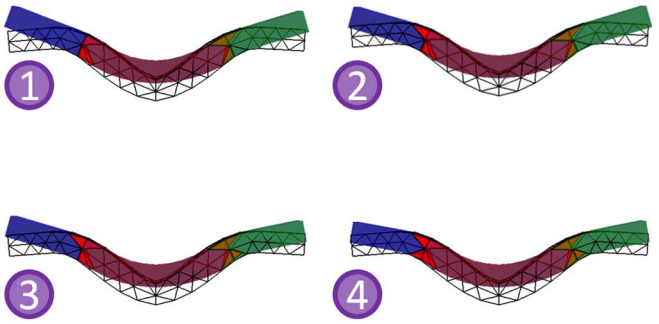


Fig. 10. Sequence of snap shots of applying postsegmentation adjustment. Each adjustment of boundary element results in a lower global error.

5.3 Automatic Seeding and Postsegmentation Adjustment

Different initial seedings could result in different final segmentations. We would like to provide an objective segmentation with as little user interference as possible.

The autoseeding algorithm starts with a single partition which will span the whole mesh. Each element is assigned an element error, e_E:

$$e_E = \frac{\|\tilde{\mathbf{u}}_E - \mathbf{u}_E\|^2}{V_E}, \tag{14}$$

where $\tilde{\mathbf{u}}_E$ and $\mathbf{u}_E$ are 12×1 vectors representing the approximated and target displacements of the tetrahedral element. V_E is the volume of the element. The element with highest e_E indicates a location that the approximation of current segmentation has the largest deviation from the target. Naturally, it is to be chosen as the seed for a new partition. After the new seeding is selected, we resegment the whole mesh using Algorithm 5.1. This process continues until the desired number of partitions is reached or the approximation error is reduced to a certain level which can be evaluated in terms of global error

$$e_G = \frac{\|\tilde{\mathbf{u}} - \mathbf{u}\|^2}{V}, \tag{15}$$

where V is the volume of the mesh at rest shape. In our experiment, we set this threshold as 5 percent e_G. Fig. 9 shows the step-by-step snap shots of our autoseeding algorithm running on the bar model with additional user-specified constraint nodes placed on the two ends of the bar.

When segmentation is finished, extra refinement can be performed in order to tune the resulting segmentation for a lower global error. The postsegmentation adjustment is straightforward: for each element connected by multiple partitions, we compute e'_G as if it were moved to its neighboring partition. Because the boundary DOFs of the partitions are changed due to the movement of the element, e'_G normally does not equal to e_G. If $e'_G < e_G$, which indicates an error-decreasing adjustment, the element swap is accepted. Fig. 10 shows the step-by-step results of applying postsegmentation adjustment to a segmentation. The approximated deformation fits the target better after each adjustment.

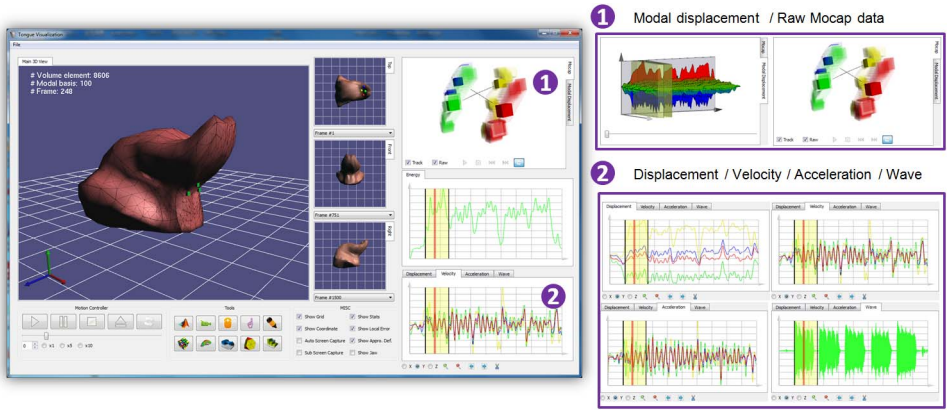


Fig. 11. A snapshot of the user interface of the proposed system.

6 Experiment Results and Analysis

Experiments have been conducted in order to demonstrate promising potential of the proposed framework with the evaluation and feedbacks from the domain experts, including a group of certified speech-language pathologists (two of them are the coauthors Drs. Vick and Campbell). Additional equipped features such as strain energy-based visualization and shape morphing are presented in Sections 6.3 and 6.4, respectively.

6.1 Interface

Fig. 11 provides an overview of the interface of our framework. Next to the main 3D view, we embed three small grided auxiliary views with three different orthogonal perspectives, top, front, and lateral, respectively. The raw mocap data as well as the displacement, velocity, and acceleration at each mocap sensor in the x, y, and z directions are also visualized in the right portion of the interface. The modal displacement is represented as a 3D surface with color map in a tabbed view. In addition, the acoustic signal and associated wave file can also be visualized and played. All the views are fully linked such that user's input in one view is reflected in the rest of the views.

6.2 Synchronized Motion Replay and Evaluation

The animation of the tongue deformation is tightly coupled with all the auxiliary views and synchronized with the experimentally recorded sound wave. This visualization strategy provides impressive and comprehensive awareness of the tongue movement. The domain experts confirmed the good potential and facilitation of the framework for their speech clinics and researches.

Although we have a very sparse mocap sensor placement over the tongue, the deformation simulated with the proposed framework is informative. Fig. 12 shows the sequences of snapshots of the 3D tongue shape from four typical CV syllable productions: /k/ is a *voiceless velar plosive* consonantal sound which is produced by obstructing the airflow in the vocal tract. It is articulated mostly with the back of the tongue; /r/, on the other hand is a common voiced *palato-alveolar affricate* which is articulated with the blade of the tongue behind the alveolar ridge forming domed curly blade; /t/ is a *voiceless dental plosive* while /l/ is a type of *alveolar lateral approximant*.

Due to the simplified simulation and geometric deviation of the tongue among subjects, the simulated shapes could differ from the real cases (for instance, /la/ in Fig. 12 has an overcurved tip rise than normal). Unfortunately, quantitative measurement over the movement of the tongue is rather rare in previous works. Furthermore as reported in [23], even for the Speech-Language Pathologists (SLPs), the knowledge of the tongue placement is quite limited. Accordingly, a user study was arranged with domain experts for a general qualitative evaluation of the simulated shapes. Five certified SLPs have participated in. We recorded the simulated tongue motions corresponding to each individual CV train (five repeated speech sound productions) without any additional information, (e.g., no audio, 2D energy curves). There were totally 40 animation clips corresponding to four CV syllables, (e.g., /ka/, /ta/, /la/ and /ra/) from both native and nonnative speakers. 20 clips of /ka/ and /ta/ were put as one group. The rest 20 clips of /la/ and /ra/ made up the other group. The evaluation was performed pairwise at each clips group. The average percentage of correct identification was 64 percent for /ka/-/ta/ group and 100 percent for /la/-/ra/ group, respectively. Our approach is able to produce a well-estimated real-time 3D motion imaging. This largely

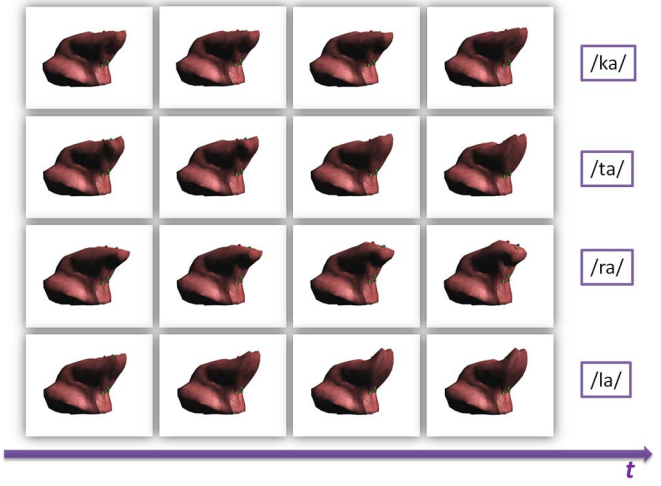


Fig. 12. Tongue deformation snapshots corresponding to different CV syllables at 20th, 40th, 60th, and 80th frame in the motion.

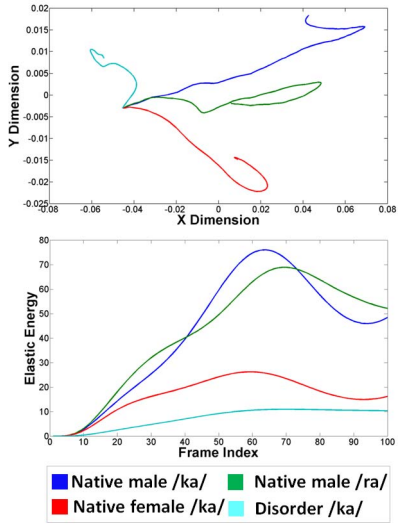


Fig. 13. Top: MDS computed with iterative majorization algorithm generates unpleasant visualization as the sequential information is lost. Bottom: energy-based curves provide better visual identification of the deformation sequences that are aligned along the time axis.

compensates for the SLPs' poor knowledge of contact along lateral margin [23].

The accuracy for /ka/ versus /ta/ is likely due to the placement of the mocap sensors that drove the visualization. The consonant /k/ is produced by raising the dorsum of the tongue to the soft palate in order to occlude the airflow and our sensor placement did not capture this motion. For this reason, the visualization of /t/ (produced by raising the tongue tip to the hard palate) was easily confused with the visualization for /k/. Additional mocap sensor(s) further back on the tongue surface could contribute a better result.

6.3 Energy-Based Low-Dimensional Visualization

Although the tongue animation delivers enriched information to the users, the animation-based visualization has fundamental limitations for lateral visual comparisons [21], [30]. A mapped low-dimensional visualization with *Multidimensional Scaling* (MDS) [16] is a commonly used strategy to handle this problem [6]. However, MDS normally exhibits poor performance in visualizing sequential motion data. As shown in Fig. 13 top, we map four motion sequences to a 2D domain using MDS (with the *iterative majorization* algorithm) and connect every two consecutive frame with a line segment. The two dimensions for the low-dimensional domain are essentially meaningless and the user can hardly tell the relative deviations among four motions at a given frame. Alternatively, we use a straightforward dimension-reduction method in our framework with 2D energy plots. The elastic energy used as a pseudoshape metric for the deformed shapes can be fast computed through (6). Consequently, each individual motion sequence becomes a 2D energy-time curve aligned along the time axis (Fig. 13 bottom).

The variation of the elastic energy curves also does a good job in motion-level visual clustering to distinguish different speech sounds. Fig. 14 displays nine curves which correspond to three motions associated with CV syllables /

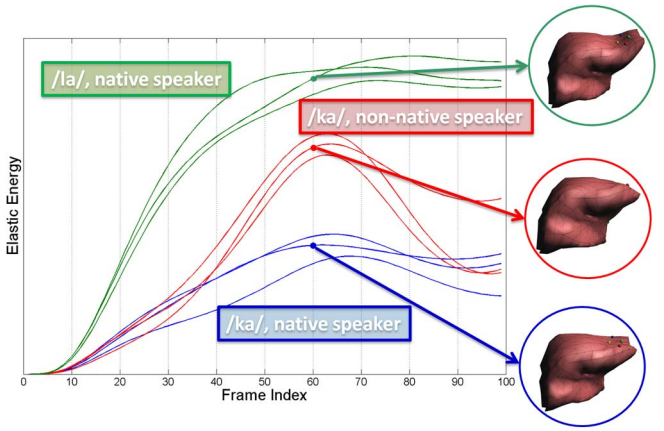


Fig. 14. 2D energy-time curves greatly facilitate the comparative visualization of multiple motion.

la/ and /ka/ from a native English speaker and a nonnative speaker, respectively. Even for the user with less related experience, it is easy to visually group these nine curves into three clusters with the help of the energy curves: the deformation energy for /la/ keeps going up during the entire production while curves corresponding to /ka/ from the native speaker has an obvious peak at the intermediate portion of the production, which could be related with the airflow break in the vocal tract when the tongue muscle is stressed in order to hold the airflow. /ka/ from the nonnative speaker has a weaker peak suggesting a different speech habit.

6.4 Morphing between Deformed Shapes

Due to the occlusion among multiple 3D geometries, effectively conveying the 3D shape difference is a challenging visualization. This problem is partially resolved in our framework by a simple yet effective morphing algorithm. As linearly blending two rotational deformations does not yield a valid intermediate rotation, most existing morphing techniques devote special efforts to handling the rotation components [1], [17], [39]. In this framework, we take advantage of the linear deformable model and only interpolate the linear part of the deformation that can be represented with modal displacement. Assume the linear modal displacements for source and target deformations are $\mathbf{q}_0$ and $\mathbf{q}_1$, respectively. An intermediate linear modal deformation, $\mathbf{q}_t$ can be simply determined by $\mathbf{q}(t) = t\mathbf{q}_0 + (1-t)\mathbf{q}_1, 0 \leq t \leq 1$, where t is the linear blending parameter. Similar to modal warping, an appropriate rotation is further applied to the blended linear deformation afterwards. The color map is also employed to guide the user's attention toward the regions where larger shape change occurs. The difference between the source and target deformations for each element is computed with (14) and the elements with high difference are colored with red and the elements with low difference are colored with blue. We linearize the *cumulative distribution function* of the element differences in order to make an even distribution of the color over the mesh.

Fig. 15 is an example of the proposed morphing algorithm. The source deformation of the tongue gradually morphs to the target deformation. With the axillary views,

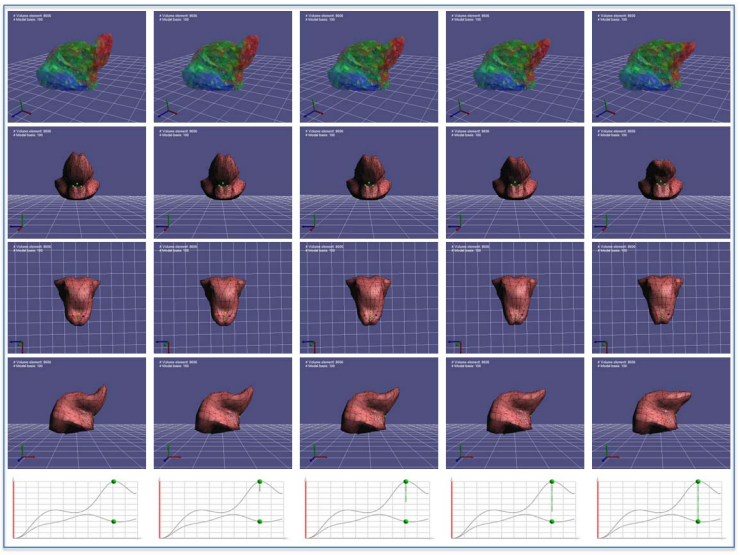


Fig. 15. Snap shots of the proposed morphing algorithm: the source and target deformations are two individual frames from the motions of /la/ and /ka/, respectively. The morphing provides an intuitive visual awareness of the 3D shape difference with a occlusion-free style.

the users are also able to observe the morphing from different angles simultaneously as shown in Fig. 15 while directly displaying the source and target deformations does not deliver intuitive information of the shape difference because of occlusion (Fig. 16).

6.5 Deformation Segmentation

The proposed segmentation algorithm highlights the deformation pattern for the given shape. Fig. 17 shows the results of our deformation segmentation algorithm (Algorithm 5.1). The target deformation is set as the single frame with the highest elastic energy from a motion of /ta/. The first row in the figure consists of intermediate results of the segmentation with rigid modes while the second row is with the constraint modes. Three initial seedings are chosen corresponding to the tongue tip and the two locations with anchor nodes. The segmentation with constraint modes approximates the target deformation with seamless interface attachment and is preferred.

Fig. 18 shows the results of the segmentation targeting the deformation sequences from a native English speaker (top row) and a speaker who stutters (bottom row). The proposed automatic segmentation method is used so that no initial seedings are specified and the threshold is set as 5 percent e_G, i.e., the segmentation stops when the approximation error over the mesh is smaller than 5 percent of shape difference between the approximated shape and the target deformation. From the top row, which is the results of the native speaker, it can be seen that /la/ requires a salient deformation at the tip and blade parts of the tongue. /ka/ yields a more regular deformation pattern such that only two partitions are sufficient to approximate it with very small error. The deformation pattern of /ra/ is somewhat similar to /la/ at the tip part. However, /ra/ has a much stronger mid-back tongue deformation. On the other hand if we look at the segmentation results of the speaker with the communication disorder as shown in the bottom row in Fig. 18, a completely different deformation pattern is exhibited. For CVs /ra/ and /la/ which need more tip movement, we can see an asymmetric tip deformation and in both CVs, the right portion of the tip is more irregular so that a partition has to be assigned there. This unique deformation pattern may suggest some morbid pronunciation habit of the speaker. In /ka/, the tip deformation region is much larger than that of the native speaker. This indicates the patient may experience difficulty with the subtle control of the tongue tip. While the domain experts confirm the potential of the proposed segmentation technique for facilitating diagnosis and treatment of the patients with communication disorder, a larger sample of individuals with disorders would need to be gathered in order to draw conclusions. The utility of this visualization technique in future studies will help examine and characterize the speech motor control of disordered populations.

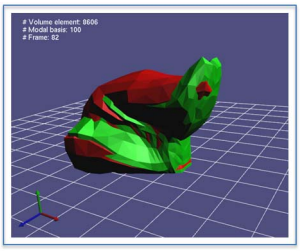


Fig. 16. Occlusion prevents intuitive visualization of shape difference.

6.6 Time Performance

Our framework is implemented using Microsoft Visual C++ 2010 on a Windows 7 PC with Intel Core2 Quad 2.4 GHz CPU, 8 GB DDR2 RAM, and NVIDIA GeForce GTX 8800 GPU with 768 MB DDR3 VRAM. GPU acceleration [7], [45] is used in the simulation of deformations based on the normalized mocap data, which gains a threefold boost in terms of Frames Per Second (FPS). With 100 modal bases, the simulation runtime FPS is 36.3. Real-time simulated tongue shape could be very useful for many applications such as language training, where vocal misbehaviors can be timely identified and corrected. Table 1 shows the detailed time performance for the techniques used in this framework.

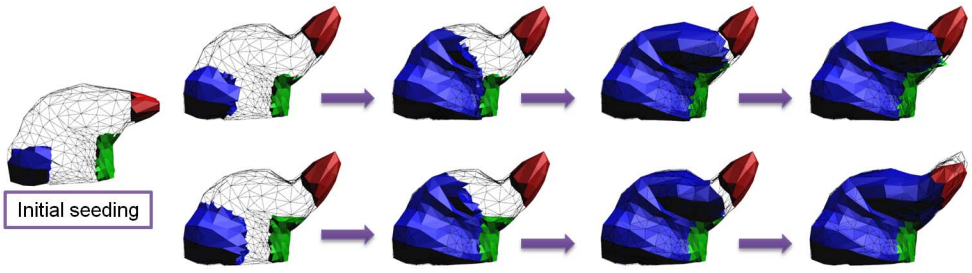


Fig. 17. Segmentation of /ta/ with rigid modes (top) and constraint modes (bottom).

7 Conclusion and Future Work

In this framework, we start from the mocap data collected during human speech production, reconstruct and visualize the 3D volumetric deformation of the human tongue with constrained modal warping. In addition, several techniques are presented to further explore the modeled 3D deformation including deformation segmentation, energy-based comparative visualization and morphing. Our experimental results confirm the promising functionality and effectiveness of our framework.

However, there are still many limitations in the current stage of this work. First, the placement of the mocap sensors is very sparse at the blade portion of the tongue. It is because the dorsal part of the tongue is hardly reachable and very difficult to be attached with accurate position specification. As a result, *dorsal consonants* may not able to be accurately simulated with our framework as these types of consonants (e.g., /g/, /y/ or /w/) are articulated with the mid body of the tongue. Our evaluation also confirms with this shortcoming. To facilitate the lateral shape comparison across subjects or different speech sounds, we did not use subject-specified mesh. The deviation between the real geometry of the subject' tongue and the mesh used for modeling could also lead to unnatural results. The proposed simulation has made a significant simplification of the elastic model because all the visualization and the following analysis are created from only five mocap sensors. The advantage of such system configuration is obvious: it is cost-efficient; the equipment set-up is straightforward; and the experiment can be repeated easily while the result is qualitative correct and visually intuitive. However, how such simplification affects the final accuracy of the simulation remains largely unknown to us, as we lack a mechanism for a quantitative evaluation of the simulated deformation. This raises several possible follow-up studies on this topic, e.g., How to place the limited number of sensors to obtain the best simulation result or should the placement of the sensors differ for different speech production? In the future, we will incorporate the analysis from different source data and make a comprehensive quantitative evaluation.

Speech production is a complex biomechanical procedure involving the coupling of jaw, tongue, velum, vocal tract, hyoid and many other related bony and muscular structures [46]. A more complex model may be more appropriate for speech analysis including the interaction of these components [35]. Incorporating the modeling of a real communication scenario is another very challenging but interesting direction for future research where a user-specific tongue model is more suitable than normalizing the mocap data to a uniform model. In addition, some derived parameters are also quite important based on the simulated deformation, such as the estimation of the muscle activation [32], which is difficult to measure directly with regular equipment.

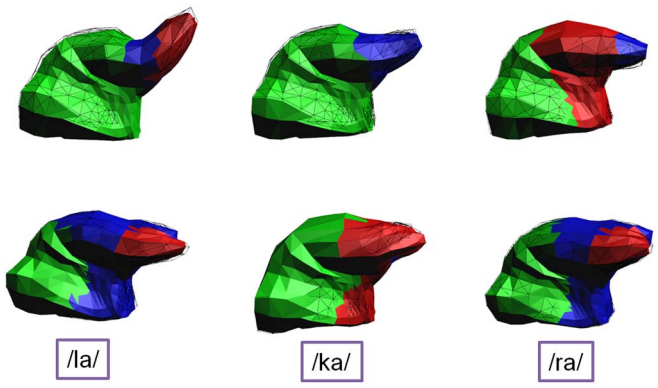


Fig. 18. Results of the automatic segmentation of the tongue of a native English speaker (top) and a speaker with communication disorder (bottom).

TABLE 1
Time Performance

Method	Time performance
Simulation with GPU	36.3 FPS
Energy-based Morphing	157.5 FPS
Segmentation with rigid modes	40.7 *sec.* on avg.
Segmentation with constraint modes	8.3 *min.* on avg.

Acknowledgments

The authors would like to thank anonymous reviewers for their constructive comments and suggestions. Yin Yang and Xiaohu Guo are partially supported by the NSF of USA under Grants Nos. CNS-1012975 and IIS-1149737.

References

[1] M. Alexa, D. Cohen-Or, and D. Levin, "As-Rigid-as-Possible Shape Interpolation," *Proc. SIGGRAPH Conf.*, pp. 157-164, 2000.
[2] P. Badin, G. Bailly, L. Revéret, M. Baciu, C. Segebarth, and C. Savariaux, "Three-Dimensional Linear Articulatory Modeling of Tongue, Lips and Face, Based on MRI and Video Images," *J. Phonetics*, vol. 30, no. 3, pp. 533-553, 2002.
[3] T. Baer, J. Gore, S. Boyce, and P. Nye, "Application of MRI to the Analysis of Speech Production," *Magnetic Resonance Imaging*, vol. 5, no. 1, pp. 1-7, 1987.
[4] Y. Bao, X. Guo, and H. Qin, "Physically Based Morphing of Point-Sampled Surfaces," *Computer Animation and Virtual Worlds*, vol. 16, pp. 509-518, July 2005.

[5] S. Buchaillard, P. Perrier, and Y. Payan, "A Biomechanical Model of Cardinal Vowel Production: Muscle Activations and the Impact of Gravity on Tongue Positioning," *The J. the Acoustical Soc. of Am.,* vol. 126, no. 4, pp. 2033-2051, 2009.

[6] W. Chen, Z. Ding, S. Zhang, A. MacKay-Brandt, S. Correia, H. Qu, J.A. Crow, D.F. Tate, Z. Yan, and Q. Peng, "A Novel Interface for Interactive Exploration of DTI Fibers," *IEEE Trans. Visualization and Computer Graphics,* vol. 15, no. 6, pp. 1433-1440, Nov./Dec. 2009.

[7] M.G. Choi and H.-S. Ko, "Modal Warping: Real-Time Simulation of Large Rotational Deformation and Manipulation," *IEEE Trans. Visualization and Computer Graphics,* vol. 11, no. 1, pp. 91-101, Jan. 2005.

[8] R. Craig, *Structural Dynamics: An Introduction to Computer Methods.* Wiley, 1981.

[9] S.F. Els, F.L.V. Ogt, K.V.D. Doel, J.E. Lloyd, and O. Guenther, "Artisynth: An Extensible, Cross-Platform 3D Articulatory Speech Synthesizer," *Proc. Conf. Auditory and Visual Speech Processing,* 2005.

[10] Q. Fang, S. Fujita, X. Lu, and J. Dang, "A Model-Based Investigation of Activations of the Tongue Muscles in Vowel Production," *Acoustical Science and Technology,* vol. 30, no. 4, pp. 277-287, 2009.

[11] Y. Fung and P. Tong, *Classical and Computational Solid Mechanics,* Advanced Series in Engineering Science. World Scientific, 2001.

[12] J.-M. Gérard, J. Ohayon, V. Luboz, P. Perrier, and Y. Payan, "Indentation for Estimating the Human Tongue Soft Tissues Constitutive Law: Application to a 3D Biomechanical Model," *Medical Simulation,* vol. 3078, pp. 77-83, 2004.

[13] J.-M. Gérard, P. Perrier, and Y. Payan, "3D Biomechanical Tongue Modeling to Study Speech Production," *Proc. Speech Production: Models, Phonetic Processes, and Techniques,* pp. 85-102, 2006.

[14] J.-M. Gérard, R. Wilhelms Tricarico, P. Perrier, and Y. Payan, "A 3D Dynamical Biomechanical Tongue Model to Study Speech Motor Control," *Research Developments in Biomechanics,* vol. 1, pp. 49-64, 2003.

[15] J.R. Green, C.A. Moore, M. Higashikawa, and R.W. Steeve, "The Physiologic Development of Speech Motor Control: Lip and Jaw Coordination," *J. Speech, Language, and Hearing Research,* vol. 43, no. 1, pp. 239-255, 2000.

[16] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning.* Springer, 2003.

[17] S.-M. Hu, C.-F. Li, and H. Zhang, "Actual Morphing: A Physics-Based Approach to Blending," *Proc. Ninth ACM Symp. Solid Modeling and Applications,* pp. 309-314, 2004.

[18] J. Huang, Y. Tong, K. Zhou, H. Bao, and M. Desbrun, "Interactive Shape Interpolation through Controllable Dynamic Deformation," *IEEE Trans. Visualization and Computer Graphics,* vol. 17, no. 7, pp. 983-992, July 2011.

[19] Q.-X. Huang, M. Wicke, B. Adams, and L. Guibas, "Shape Decomposition Using Modal Analysis," *Computer Graphics Forum,* vol. 28, no. 2, pp. 407-416, 2009.

[20] T. Hughes, *The Finite Element Method: Linear Static and Dynamic Finite Element Analysis.* Dover Publications, 2000.

[21] D. Keefe, M. Ewert, W. Ribarsky, and R. Chang, "Interactive Coordinated Multiple-View Visualization of Biomechanical Motion Data," *IEEE Trans. Visualization and Computer Graphics,* vol. 15, no. 6, pp. 1383-1390, Nov./Dec. 2009.

[22] S.A. King and R.E. Parent, "A 3D Parametric Tongue Model for Animated Speech," *The J. Visualization and Computer Animation,* vol. 12, no. 3, pp. 107-115, 2001.

[23] S. Mcleod, "Speechclanguage Pathologists Knowledge of Tongue/Palate Contact for Consonants," *Clinical Linguistics and Phonetics,* vol. 25, nos. 11/12, pp. 1004-1013, 2011.

[24] W. Moss, M.C. Lin, and D. Manocha, "Constraint-Based Motion Synthesis for Deformable Models," *Computer Animation and Virtual Worlds,* vol. 19, nos. 3/4, pp. 421-431, 2008.

[25] M. Müller, J. Dorsey, L. McMillan, R. Jagnow, and B. Cutler, "Stable Real-Time Deformations," *Proc. ACM SIGGRAPH/Eurographics Symp. Computer Animation,* pp. 49-54, 2002.

[26] Northern Digital Inc., "Wave Speech Research System," http://www.fermentas.com/techinfo/nucleicacids/maplambda.htm, 2012.

[27] D.J. Ostry, E. Vatikiotis-Bateson, and P.L. Gribble, "An Examination of the Degrees of Freedom of Human Jaw Motion in Speech and Mastication," *J. Speech, Language, and Hearing Research,* vol. 40, no. 6, pp. 1341-1351, 1997.

[28] A. Pentland and J. Williams, "Good Vibrations: Modal Dynamics for Graphics and Animation," *Computer Graphics,* vol. 23, no. 3, pp. 207-214, 1989.

[29] E. Pernkopf, *Atlas of Topographical and Applied Human Anatomy: Head and Neck,* third ed. Williams Wilkins, Dec. 1989.

[30] G. Robertson, R. Fernandez, D. Fisher, B. Lee, and J. Stasko, "Effectiveness of Animation in Trend Visualization," *IEEE Trans. Visualization and Computer Graphics,* vol. 14, no. 6, pp. 1325-1332, Nov. 2008.

[31] A. Shamir, "A Survey on Mesh Segmentation Techniques," *Computer Graphics Forum,* vol. 27, no. 6, pp. 1539-1556, 2008.

[32] E. Sifakis, I. Neverov, and R. Fedkiw, "Automatic Determination of Facial Muscle Activations from Sparse Motion Capture Marker Data," *ACM Trans. Graphics,* vol. 24, no. 3, pp. 417-425, July 2005.

[33] B.C. Sonies, T.H. Shawker, T.E. Hall, L.H. Gerber, and S.B. Leighton, "Ultrasonic Visualization of Tongue Motion During Speech," *The J. the Acoustical Soc. of Am.,* vol. 70, no. 3, pp. 683-686, 1981.

[34] I. Stavness, A.G. Hannam, J.E. Lloyd, and S. Fels, "Predicting Muscle Patterns for Hemimandibulectomy Models," *Computer Methods in Biomechanics and Biomedical Eng.,* vol. 13, no. 4, pp. 483-491, 2010.

[35] I. Stavness, J.E. Lloyd, Y. Payan, and S. Fels, "Coupled Hard-Soft Tissue Simulation with Contact and Constraints Applied to Jaw-Tongue-Hyoid Dynamics," *Int'l J. for Numerical Methods in Biomedical Eng.,* vol. 27, no. 3, pp. 367-390, 2011.

[36] M. Stone, "A Three-Dimensional Model of Tongue Movement Based on Ultrasound and X-Ray Microbeam Data," *The J. the Acoustical Soc. of Am.,* vol. 87, no. 5, pp. 2207-2217, 1990.

[37] M. Stone, E.P. Davis, A.S. Douglas, M.N. Aiver, R. Gullapalli, W.S. Levine, and A.J. Lundberg, "Modeling Tongue Surface Contours from Cine-MRI Images," *J. Speech, Language, and Hearing Research,* vol. 44, no. 5, pp. 1026-1040, 2001.

[38] M. Stone and A. Lundberg, "Three-Dimensional Tongue Surface Shapes of English Consonants and Vowels," *The J. the Acoustical Soc. of Am.,* vol. 99, no. 6, pp. 3728-3737, 1996.

[39] R.W. Sumner, M. Zwicker, C. Gotsman, and J. Popović, "Mesh-Based Inverse Kinematics," *Proc. SIGGRAPH '05,* pp. 488-495, 2005.

[40] H. Takemoto, "Morphological Analyses of the Human Tongue Musculature for Three-Dimensional Modeling," *J. Speech, Language, and Hearing Research,* vol. 44, no. 1, pp. 95-107, 2001.

[41] F. Vogt, J. Lloyd, S. Buchaillard, P. Perrier, M. Chabanas, Y. Payan, and S. Fels, "Efficient 3D Finite Element Modeling of a Muscle-Activated Tongue," *Biomedical Simulation,* vol. 4072, pp. 19-28, 2006.

[42] R. Wilhelms-Tricarico, "Physiological Modeling of Speech Production: Methods for Modeling Soft-Tissue Articulators," *The J. the Acoustical Soc. of Am.,* vol. 97, no. 5, pp. 3085-3098, 1995.

[43] A. Witkin and W. Welch, "Fast Animation and Control of Nonrigid Structures," *Proc. SIGGRAPH '90,* pp. 243-252, 1990.

[44] H.-B. Yan, S.-M. Hu, and R. Martin, "Morphing Based on Strain Field Interpolation," *Computer Animation and Virtual Worlds,* vol. 15, nos. 3/4, pp. 443-452, July 2004.

[45] Y. Yang, G. Rong, L. Torres, and X. Guo, "Real-Time Hybrid Solid Simulation: Spectral Unification of Deformable and Rigid Materials," *Computer Animation and Virtual Worlds,* vol. 21, nos. 3/4, pp. 151-159, 2010.

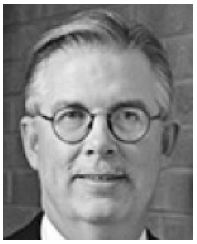
[46] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Quantitative Association of Vocal-Tract and Facial Behavior," *Speech Comm.,* vol. 26, pp. 23-43, Oct. 1998.

Yin Yang is currently working toward the PhD degree in the Department of Computer Science, The University of Texas at Dallas, where he is a research assistant. His research focuses on physics-based animation/simulation and related applications, scientific visualization, and medical imaging. He is a student member of the IEEE and the IEEE Computer Society.

Xiaohu Guo received the PhD degree in computer science from the State University of New York at Stony Brook in 2006. He is an associate professor of computer science at the University of Texas at Dallas. His research interests include computer graphics, animation, and visualization, with an emphasis on geometric, and physics-based modeling. His current researches at UT-Dallas include: spectral geometric analysis, deformable models, centroidal Voronoi tessellation, GPU algorithms, 3D and 4D medical image analysis, etc. He received the prestigious National Science Foundation CAREER Award in 2012. He is a member of the IEEE and the IEEE Computer Society. For more information, please visit http://www.utdallas.edu/xguo.

Jennell Vick received the PhD degree from the University of Washington in Seattle, in 2008, where she studied developmental and disordered speech motor control under the direction of Dr. Christopher Moore. She is a CCC-SLP holder. Her postdoctoral work in speech motor control in disordered populations and advanced statistical modeling was completed at The University of Texas at Dallas Callier Center for Communication Disorders under the direction of Dr. Thomas Campbell. She is now an assistant professor in the Departments of Psychology, Biomedical Engineering, and Pediatrics at Case Western Reserve University.

Luis G. Torres received the BS degree in computer science in 2010 at the University of Texas at Dallas, where he conducted research in physically based modeling. Currently, he is working toward the PhD degree in computer science at the University of North Carolina at Chapel Hill, where he conducts research in medical robotics and motion planning algorithms. He is a recipient of the National Science Foundation Graduate Research Fellowship. He is a student member of the IEEE.



Thomas F. Campbell received the bachelor of science degree in education, the master of arts degree in speech-language pathology from The University of Nebraska-Lincoln, and the PhD degree in communicative disorders from The University of Wisconsin-Madison. He is the Sara T. Martineau endowed professor in communication disorders in the School of Behavioral and Brain Sciences at The University of Texas at Dallas and the executive director of the UTD Callier Center for Communication Disorders. Throughout his career, he has worked as a speech-language pathologist, clinical educator, and federally funded researcher. He has conducted several NIH funded investigations including longitudinal studies of speech and language skills in children recovering from severe traumatic brain injury. His current research focuses on the identification of physiological, environmental and genetic variables for the early identification of speech and language disorders in children with developmental and acquired communication disorders. He is a trustee of the American Speech-Language-Hearing Foundation.