Speech animation using electromagnetic articulography as motion capture data

Ingmar Steiner [1,2], *Korin Richmond* [3], *Slim Ouni* [4]

[1]Multimodal Computing and Interaction, Saarland University, Germany;
[2]DFKI GmbH, Saarbrücken, Germany;
[3]CSTR, University of Edinburgh, UK;
[4]Université de Lorraine, LORIA, UMR 7503, France
`ingmar.steiner@dfki.de`; `korin@cstr.ed.ac.uk`; `slim.ouni@loria.fr`

Abstract

Electromagnetic articulography (EMA) captures the position and orientation of a number of markers, attached to the articulators, during speech. As such, it performs the same function for speech that conventional motion capture does for full-body movements acquired with optical modalities, a long-time staple technique of the animation industry.

In this paper, EMA data is processed from a motion-capture perspective and applied to the visualization of an existing multimodal corpus of articulatory data, creating a kinematic 3D model of the tongue and teeth by adapting a conventional motion capture based animation paradigm. This is accomplished using off-the-shelf, open-source software. Such an animated model can then be easily integrated into multimedia applications as a digital asset, allowing the analysis of speech production in an intuitive and accessible manner.

The processing of the EMA data, its co-registration with 3D data from vocal tract magnetic resonance imaging (MRI) and dental scans, and the modeling workflow are presented in detail, and several issues discussed.

Index Terms: speech production, articulatory data, electromagnetic articulography, vocal tract, motion capture, visualization

1. Introduction

Speech scientists have a number of medical imaging modalities at their disposal to capture hidden articulatory motion during speech, including realtime magnetic resonance imaging (MRI), ultrasound tongue imaging, and electromagnetic articulography (EMA) [1], the latter more recently in 3D [2]. Such techniques are commonly used to analyze and visualize the articulatory motion of human speakers. Indeed, the resulting data has been applied to articulatory animation for audiovisual (AV) speech synthesis [3–8]; using motion-capture data to animate such models can lead to significant improvements over rule-based animation [9]. However, these synthesizers are generally focused towards clinical applications such as speech therapy or biomechanical simulation.

While the lip movements can be acquired using optical tracking and the teeth and jaw are rigid bodies, the tongue is more complex to model, since its anatomical structure makes it highly flexible and deformable. The majority of previous work has modeled the tongue based on static shapes (obtained from MRI) and statistical parametric approaches to deforming them by vertex animation [6, 7] or finite element modeling [10–14]. [4] and [15] are an exception in that they deform

the tongue model using a skeletal animation technique, which rigs the tongue mesh on an armature of pseudo-bones exerting varying degrees of influence on the position of the mesh vertices.

In this paper, we describe a technique for articulatory animation, i.e., visualization of movements of the articulators during speech, by adapting a conventional motion capture based animation paradigm (Section 2). A form of skeletal animation is applied to the tongue, but driven directly by EMA data, without the intermediate abstraction of independent parameters. The static 3D model is extracted from volumetric MRI and dental scans from the same speaker as the EMA data. In this way, we avoid the issue of cross-speaker vocal tract normalization.

The reader is kindly requested to note that this technique is by no means intended to provide an accurate model of tongue shapes or movements, as previous work using biomechanical models does; rather, the advantage here is the lightweight implementation, which relies exclusively on the articulatory data itself, processed as described in Section 3.2, and a conventional 3D modeling workflow using industry standards and off-the-shelf, open-source software (cf. Section 3.3); this eliminates the implementation issues which burden more ambitious frameworks [13, for instance].

2. EMA as motion capture

Motion capture has been used for many years to analyze human movements, gait and gestures, and to control the motions and expressions of avatars and virtual characters in a variety of media [16]. Accordingly, an entire industry has formed around the acquisition and processing of motion capture data, and the rigging and animation of 3D models that are controlled using this data. There is a huge community of producers and consumers of motion capture, countless databases and stock animation resources, and a rich ecosystem of proprietary and open-source software tools to manipulate them.

Despite attempts to design and promote elaborate and flexible universal standards for motion capture data [17,18], certain other industry-backed file formats have evolved to become de-facto standards instead [19]. One format in particular, Biovision Hierarchy (BVH), has survived the company that created it and is now widely supported, presumably because it is simple and clearly defined, straightforward to implement, and human-readable [20,21].

A BVH file contains ASCII text divided into two sections, which define the skeletal objects underlying the motion cap-

ture data in a `HIERARCHY` of joints, and the `MOTION` of these objects, respectively. Although most BVH data describes the shape and motion of bipedal humanoids, this is by no means a requirement; multiple independent objects of arbitrary internal structure can be defined.

This feature of the BVH format, along with its widespread support, is a simple, but fundamental point for the premise of the present paper. By interpreting the coils of EMA data as objects whose motion is captured over time and describing it in a format such as BVH, it becomes straightforward to introduce the EMA data into a conventional motion capture processing workflow and to use it to control the visualization of geometric models of speech articulation.

3. Articulatory animation

Based on previous work exploring the feasibility of articulatory animation based on EMA in a motion capture paradigm [22], we apply our articulatory modeling approach to a large multimodal corpus of articulatory data.

While we had previously experimented with a layout of seven EMA coils on the tongue, we found that the resulting animation was vulnerable to noise and errors in the seven-coil data, potentially due to the influence of any or all of the following factors: (a) coil detachments during the recording session, (b) faulty coil hardware, (c) interference due to coil proximity, and (d) issues with the post-processing software [23].

3.1. Multimodal corpus

The `mngu0` corpus [24] contains articulatory data obtained from a single male speaker of British English, using a variety of modalities. This data is freely available for research purposes. Several hours of speech were recorded using a Carstens AG500 articulograph and video camera, yielding 3D EMA data and synchronized video and audio for more than 2 000 utterances [25]. In addition, a set of volumetric and dynamic MRI scans were acquired of the speaker's vocal tract during sustained and vowel-consonant-vowel speech production, respectively [26]. Finally, dental casts were made of the speaker's teeth, gums, and palate; the casts were then scanned to produce high-resolution, digital 3D models.

3.2. Data processing

3.2.1. EMA data

For this paper, the "day 1" subset of the EMA data in the `mngu0` corpus was used; the layout features EMA coils on the upper and lower lip, three tongue coils (T1-3, on the tip, blade, and dorsum, respectively), and mandibular incisors, as well as reference coils behind each ear, on the bridge of the nose, and on the maxillary incisors.

The large number of utterances and unusually high quality of the data greatly outweigh the sparse coverage of the tongue surface represented by only three coils. Nevertheless, it turns out that the spline IK based animation (cf. Section 3.3.2) yields satisfactory results with this arrangement.

The `mngu0` EMA data is provided in binary `EST_Track` format, which can be manipulated using the Edinburgh Speech Tools [27]. The data was processed directly from the raw `amp` files produced by the AG500, using a custom version of TAPADM [28]. The EMA coil orientation is encoded not as two Euler angles, as in the `pos` file format produced by the Carstens software [29], but as three rotation normals.

```
HIERARCHY
ROOT T3
{
  OFFSET  -0.083  -0.686  0.722
  CHANNELS 6 XPosition YPosition ZPosition ...
      XRotation YRotation ZRotation
  End Site
  {
    OFFSET  0.000 -1.000  0.000
  }
}
ROOT upperlip
{
  OFFSET  -0.945  0.297 -0.130
  CHANNELS 6 XPosition YPosition ZPosition ...
      XRotation YRotation ZRotation
  End Site
  {
    OFFSET  0.000 -1.000  0.000
  }
}
MOTION
Frames: 724
Frame Time: 0.005
6.790 12.273  1.200 -8.137  56.687  -1.782  ...
    -6.083  12.965  0.386 19.706  -11.455 ...
    -52.567 0.175 0.978 -3.002  3.939 24.401  ...
    -51.690 0.053 -0.487  -2.579  56.493  9.144 ...
    -2.778  0.005 -0.006  5.717 56.740  1.991 ...
    7.709 -0.003  -0.026  0.006 -57.279 -1.344  ...
    0.390 0.258 1.551 -0.986  7.964 -30.909 ...
    -47.582 0.106 3.215 0.254 -8.542  -56.293 ...
    -6.396  0.112 5.056 -0.049  -4.804  -39.311 ...
    41.405  0.096 -1.061  -0.194  -54.183 ...
    17.068  -7.465
6.789 12.272  1.201 -8.140  56.686  -1.790  ...
    -6.083  12.966  0.386 19.704  -11.460 ...
    -52.566 0.172 0.979 -3.003  3.946 24.397  ...
    -51.692 0.050 -0.489  -2.579  56.493  9.144 ...
    -2.788  0.005 -0.006  5.717 56.741  1.995 ...
    7.702 -0.003  -0.025  0.006 -57.279 -1.347  ...
    0.390 0.260 1.550 -0.986  7.962 -30.937 ...
    -47.564 0.107 3.215 0.254 -8.544  -56.295 ...
    -6.380  0.114 5.058 -0.051  -4.791  -39.304 ...
    41.413  0.094 -1.057  -0.196  -54.183 ...
    17.063  -7.476
```

Listing 1: *Excerpt of one BVH file (`mngu0_s1_0001.bvh`). Note that only two EMA coil objects (T3 and upperlip) are listed, and the motion data (shown for all 10 coils) is truncated after the second frame.*

The `EST_Track` files were converted to BVH format in the following manner. Each EMA coil is interpreted as an independent armature in the `HIERARCHY` and encoded as a separate `ROOT` object with six channels corresponding to the Cartesian coordinates and rotation normals of the coil's position and orientation, respectively. The values for each coil's `OFFSET` from the origin represent the coil distribution in the first data frame. Each armature is terminated by a tip whose `OFFSET` is given by a unit vector. The `MOTION` is simply copied from the corresponding channels in the `EST_Track` data. An example of the result of this conversion is given in Listing 1.

For downstream registration using the palate as a landmark (cf. Section 3.3.1), in lieu of a dedicated palate trace sweep, the convex hull of the tongue coil position samples was calculated as a contour intersected with the speaker's midsagittal plane, using VisArtico [30]. This reconstructed palate contour was then exported as a 3D point cloud.
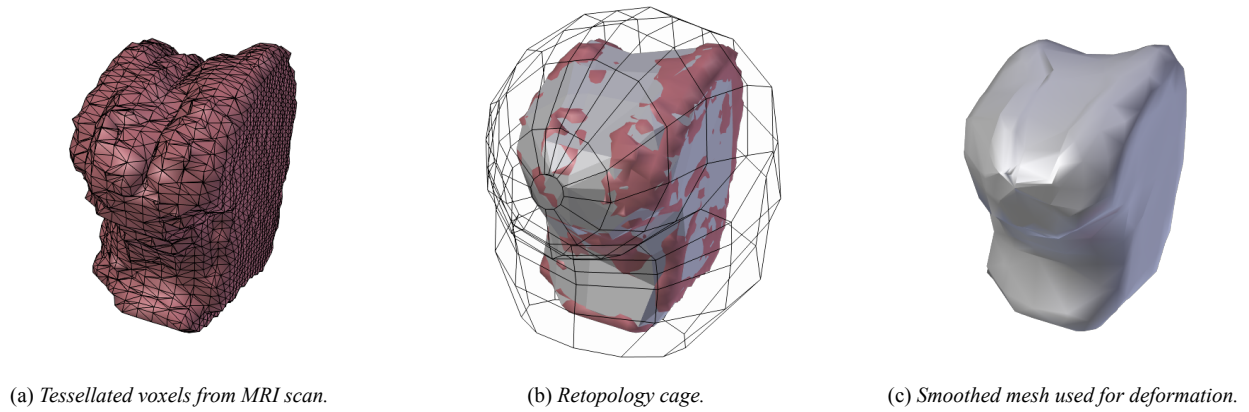
(a) *Tessellated voxels from MRI scan.*   (b) *Retopology cage.*   (c) *Smoothed mesh used for deformation.*

Figure 1: *Tongue mesh retopology.*

3.2.2. Tongue MRI data

Instead of an artificial static 3D model of the tongue and teeth, as was used for our feasibility study [22], the present paper uses a static tongue mesh extracted from the MRI subset of the mngu0 corpus. Selecting an articulatory configuration with a clearly visible cavity between the tongue and palate surfaces, the tongue from the volumetric MRI scan of the sustained vowel [ɑː] was manually segmented and exported, using OsiriX [31], which acts as a graphical user interface to a number of toolkits for manipulating and visualizing medical imaging data.

The tongue was first enclosed in a region of interest (ROI) annotation in each MRI slice, drawn by hand using a pen tablet. The pixel values outside these ROIs were then set to zero. Finally, the tongue was rendered as an isosurface within the 3D ROI formed by interpolating between the slicewise annotations, using a threshold value, and exported as a static 3D mesh (Figure 1a).

Since this mesh is generated simply by tessellating the anamorphic voxels of the volumetric MRI data, the mesh topology is extremely ill-suited to the goal of realistic deformation. As a consequence, the tongue mesh was retopologized with a simple cage, which was "shrinkwrapped" to the surface of the exported mesh (Figure 1b) and smoothed using Catmull-Clark subdivision [32] (Figure 1c); this task was performed using Blender [33].

In order to obtain a landmark to be used in subsequent registration (cf. Section 3.3.1), the hard palate was also segmented and exported as a static mesh in a manner analogous to the tongue extraction.

3.2.3. Dental scans

As the resolution of the dental scans was deemed too high for efficient processing, the vertices of the maxilla and mandible meshes were deduplicated (using MeshLab [34]) and decimated (using Blender) to roughly 5 %; this reduced the vertex count from 927 282 and 836 892 to 7 729 and 6 976 for the maxilla and mandible, respectively.

3.3. Vocal tract modeling

The MRI and dental cast scans were co-registered, and the resulting static 3D model was rigged and animated with the EMA data, as described below.
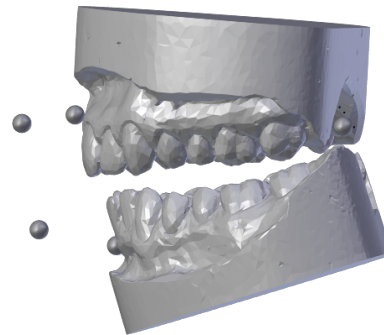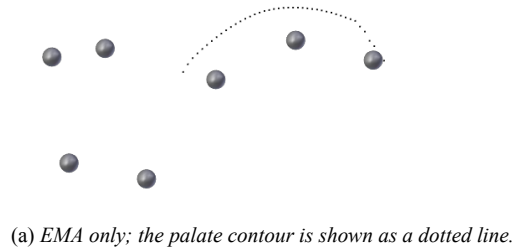

(a) *EMA only; the palate contour is shown as a dotted line.*

(b) *Dental scans added.*

Figure 2: *Registration of EMA data and dental scan meshes.*

3.3.1. Cross-modal registration

Using the palate as a landmark, the data from the different modalities was registered into the same geometric space. The palate contour from the EMA data was used to position the maxillary dental mesh, while the mandibular mesh was initially positioned accordingly, using the occlusal plane as a reference (Figure 2). The retopologized tongue mesh was co-registered with the other modalities by fitting the surface of the exported palate mesh to the palate surface of the maxillary dental scan.
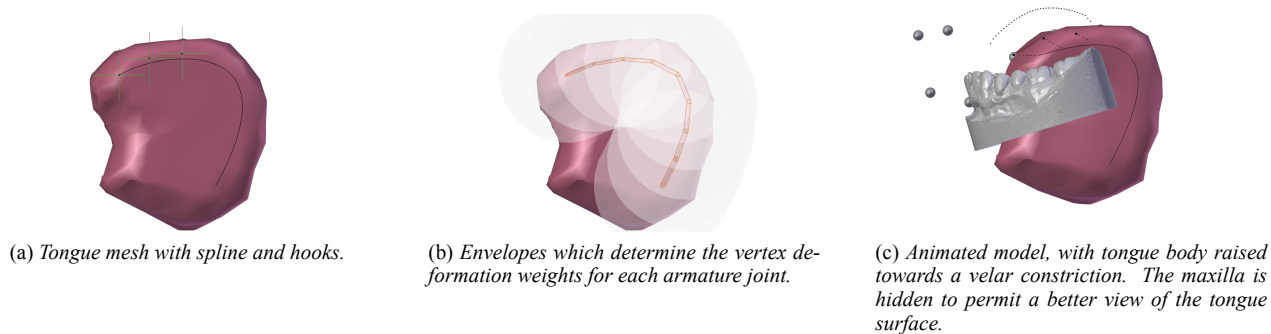
(a) *Tongue mesh with spline and hooks.*

(b) *Envelopes which determine the vertex deformation weights for each armature joint.*

(c) *Animated model, with tongue body raised towards a velar constriction. The maxilla is hidden to permit a better view of the tongue surface.*

Figure 3: *Rigging and animating the articulatory model.*

3.3.2. Rigging

A 3D non-uniform rational B-spline (NURBS) was created along, and just under, the surface of the tongue mesh, from the tongue root to the tip in the midsagittal plane. The three EMA tongue coils were then configured to act as "hooks", or control points, modifying the shape of the spline according to their deviation from the initial bind pose (Figure 3a). The hooks were then assigned as children of the respective tongue coils, so that the EMA data determines their positions.

An armature "chain" was then created and parented to the NURBS, following its shape using spline IK. The armature envelopes for this chain were expanded to enclose the upper surface and sides of the tongue mesh, and vertex groups were created for the mesh to control the tongue's deformation, with weights assigned automatically, based on the envelopes (Figure 3b).

For jaw motion, a simple modifier was added to rotate the mandibular dental mesh around a hinge, tracking the EMA coil on the mandibular incisors.

3.3.3. Animation

As a result of the rigging process described above, the EMA data drives the motion of the jaw in the articulatory model, and likewise controls the tongue by deforming the NURBS, which in turn modifies the shape of the tongue armature using spline IK. An example of the animation rendered in this manner can be seen in Figure 3c and in the supplementary material for this paper.

3.4. Evaluation

By comparing the motion of the tongue coils with that of three corresponding vertices on the tongue mesh, we receive a rough evaluation of how well the animation preserves the nature of the articulatory data. While the sparse topology of the mesh does entail a noticeable offset (cf. the example in Figure 4), the overall mean correlation of 0.95 indicates that the characteristics of the natural movements are reflected in the animation.

It should be noted that the *shape* of the tongue mesh, when it is deformed by the EMA data for a given articulatory configuration, cannot be expected to match that of static tongue shapes obtained by, e.g., sustained production in a supine posture. Nevertheless, it could be useful to compare the surfaces of such tongue shapes with those obtained through deformation of the static mesh. Since corresponding data is available in the `mngu0`
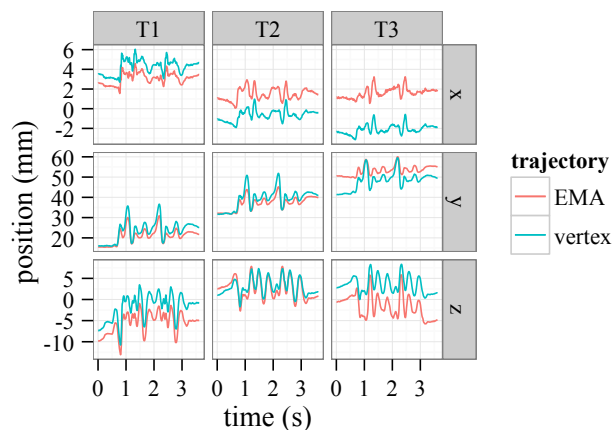

Figure 4: *EMA-measured and animated trajectories of the three tongue coils and corresponding vertices for one utterance.*

corpus, such comparisons are indeed planned in the near future. A better assessment of the tongue deformation's degree of realism would however be possible by comparing it with real-time MRI of fluent speech [35, 36]

4. Discussion and Outlook

While the kinematics of the articulatory animation appear natural, a number of issues were identified with the techniques presented here, and remain to be addressed.

- The tongue segmentation from MRI data, and subsequent retopology, represent a rather tedious manual process. Automation of some or all of the aspects of these tasks would be highly desirable.
- The manual cross-modal registration based on the palate contour is also error-prone. Actual 3D palate trace data in the EMA modality would permit a more robust surface fitting technique.
- There are currently no constraints to prohibit the 3D models from passing through one another. By integrating collision detection using a physics engine such as [37], this could be prevented, and soft body dynamics simulated.
- The placement of EMA coils determines the animation. This is both a blessing and a burden, since the lack of independent parameters could cause non-optimal coil layouts to unduly influence the animation.

- The initial bind pose in the rigging process is critical to subsequent animation; if the exact position of e.g., the jaw coil relative to the mandibular mesh can only be estimated, the overall animation will reflect a poor choice here.
- The MRI data was acquired in a supine position, but the EMA data was recorded with the speaker sitting upright. As a result, the extracted tongue mesh may well be influenced by posture and gravity [38].

In conclusion, we have presented an articulatory animation technique which is driven by articulatory data in a conventional motion capture based animation paradigm. It features a lightweight implementation using off-the-shelf, open-source software, and a footprint that is small enough to allow integration of the resulting model into applications for articulatory data exploration and real-time visualization, as well as integration into frameworks for AV speech synthesis for virtual characters, where realistic animation is more important than matching the true shape of the tongue.

Future work includes evaluating the tongue animation based on different tongue shapes or contours in real imaging data, as well as co-registration with the video data in the `mngu0` corpus, based on optical tracking of the visible EMA coils.

5. References

[1] K. M. Hiiemae and J. B. Palmer, "Tongue movements in feeding and speech," *Critical Reviews in Oral Biology & Medicine*, vol. 14, no. 6, pp. 413–429, Nov. 2003, doi:10.1177/154411130301400604.

[2] P. Hoole and A. Zierdt, "Five-dimensional articulography," in *Speech Motor Control: New developments in basic and applied research*, B. Maassen and P. van Lieshout, Eds. Oxford University Press, 2010, ch. 20, pp. 331–349.

[3] M. M. Cohen and D. W. Massaro, "Modeling coarticulation in synthetic visual speech," in *Models and Techniques in Computer Animation*, N. Magnenat-Thalmann and D. Thalmann, Eds. Springer, 1993, pp. 139–156.

[4] C. Pelachaud, C. van Overveld, and C. Seah, "Modeling and animating the human tongue during speech production," in *Computer Animation*, Geneva, Switzerland, May 1994, pp. 40–49, doi:10.1109/CA.1994.324008.

[5] S. A. King and R. E. Parent, "A 3D parametric tongue model for animated speech," *Journal of Visualization and Computer Animation*, vol. 12, no. 3, pp. 107–115, Sep. 2001, doi:10.1002/vis.249.

[6] 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," *Journal of Phonetics*, vol. 30, no. 3, pp. 533–553, Jul. 2002, doi:10.1006/jpho.2002.0166.

[7] O. Engwall, "Combining MRI, EMA & EPG measurements in a three-dimensional tongue model," *Speech Communication*, vol. 41, no. 2-3, pp. 303–329, Oct. 2003, doi:10.1016/S0167-6393(02)00132-2.

[8] S. Fagel and C. Clemens, "An articulation model for audiovisual speech synthesis: Determination, adjustment, evaluation," *Speech Communication*, vol. 44, no. 1-4, pp. 141–154, Oct. 2004, doi:10.1016/j.specom.2004.10.006.

[9] O. Engwall and P. Wik, "Real vs. rule-generated tongue movements as an audio-visual speech perception support," in *FONETIK*, Stockholm, Sweden, May 2009, pp. 30–35.

[10] M. Stone, E. P. Davis, A. S. Douglas, M. NessAiver, R. Gullapalli, W. S. Levine, and A. Lundberg, "Modeling the motion of the internal tongue from tagged cine-MRI images," *Journal of the Acoustical Society of America*, vol. 109, no. 6, pp. 2974–2982, Jun. 2001, doi:10.1121/1.1344163.

[11] J.-M. Gerard, R. Wilhelms-Tricarico, P. Perrier, and Y. Payan, "A 3D dynamical biomechanical tongue model to study speech motor control," 2003, submitted to Recent Research Developments in Biomechanics. [Online]. Available: http://arxiv.org/abs/physics/0606148

[12] J. Dang and K. Honda, "Construction and control of a physiological articulatory model," *Journal of the Acoustical Society of America*, vol. 115, no. 2, pp. 853–870, Feb. 2004, doi:10.1121/1.1639325.

[13] F. Vogt, J. E. Lloyd, S. Buchaillard, P. Perrier, M. Chabanas, Y. Payan, and S. S. Fels, "Efficient 3D finite element modeling of a muscle-activated tongue," in *Biomedical Simulation*, ser. Lecture Notes in Computer Science, M. Harders and G. Székely, Eds. Springer, 2007, vol. 4072, pp. 19–28.

[14] Y. Yang, X. Guo, J. Vick, L. G. Torres, and T. Campbell, "Physics-based deformable tongue visualization," *IEEE Transactions on Visualization and Computer Graphics*, 2012, doi:10.1109/TVCG.2012.174.

[15] M. D. Ilie, C. Negrescu, and D. Stanomir, "An efficient parametric model for real-time 3D tongue skeletal animation," in *9th International Conference on Communications (COMM)*, Bucharest, Romania, Jun. 2012, pp. 129–132, doi:10.1109/ICComm.2012.6262577.

[16] A. Menache, *Understanding Motion Capture for Computer Animation and Video Games*. Academic Press, 2000.

[17] Motion Lab Systems, *The C3D File Format User Guide*, Baton Rouge, LA, USA, Jan. 2008. [Online]. Available: http://www.c3d.org/pdf/c3dformat_ug.pdf

[18] H.-S. Chung and Y. Lee, "MCML: motion capture markup language for integration of heterogeneous motion capture data," *Computer Standards & Interfaces*, vol. 26, no. 2, pp. 113–130, Mar. 2004, doi:10.1016/S0920-5489(03)00071-0.

[19] M. Meredith and S. Maddock. Motion capture file formats explained. [Online]. Available: http://www.mmeredith.staff.shef.ac.uk/fileformats/mocapff.pdf

[20] M. Gleicher. Biovision BVH. [Online]. Available: http://research.cs.wisc.edu/graphics/Courses/cs-838-1999/Jeff/BVH.html

[21] T. Howard. BVH readers and motion capture resources. [Online]. Available: http://www.cs.man.ac.uk/~toby/bvh/

[22] I. Steiner and S. Ouni, "Artimate: an articulatory animation framework for audiovisual speech synthesis," in *ISCA Workshop on Innovation and Applications in Speech Technology*, Dublin, Ireland, Mar. 2012. [Online]. Available: http://arxiv.org/abs/1203.3574

[23] M. Stella, P. Bernardini, F. Sigona, A. Stella, M. Grimaldi, and B. G. Fivela, "Numerical instabilities and three-dimensional electromagnetic articulography," *Journal of the Acoustical Society of America*, vol. 132, no. 6, pp. 3941–3949, Dec. 2012, doi:10.1121/1.4763549.

[24] The mngu0 data set. [Online]. Available: http://mngu0.org/

[25] K. Richmond, P. Hoole, and S. King, "Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus," in *Interspeech*, Florence, Italy, Aug. 2011, pp. 1505–1508. [Online]. Available: http://www.isca-speech.org/archive/interspeech_2011/i11_1505.html

[26] I. Steiner, K. Richmond, I. Marshall, and C. D. Gray, "The magnetic resonance imaging subset of the mngu0 articulatory corpus," *Journal of the Acoustical Society of America*, vol. 131, no. 2, pp. 106–111, Feb. 2012, doi:10.1121/1.3675459.

[27] S. King, A. W. Black, P. Taylor, R. Caley, and R. Clark, *Edinburgh Speech Tools Library System Documentation*, 1.2.3 ed., Centre for Speech Technology, University of Edinburgh, Jan. 2003. [Online]. Available: http://www.cstr.ed.ac.uk/projects/speech_tools/manual-1.2.0/

[28] A. Zierdt. Three-dimensional Artikulographic Position and Align Determination with MATLAB. [Online]. Available: http://wiki.ag500.net/TAPADM

[29] Carstens Medizinelektronik, *AG500 Data Format and Data Structure*, Lenglern, Germany, 2004. [Online]. Available: http://www.ag500.de/manual/ag500/AG500_ida_format.pdf

[30] S. Ouni, L. Mangeonjean, and I. Steiner, "VisArtico: a visualization tool for articulatory data," in *Interspeech*, Portland, OR, USA, Sep. 2012, pp. 1878–1881. [Online]. Available: http://www.isca-speech.org/archive/interspeech_2012/i12_1878.html

[31] A. Rosset, L. Spadola, and O. Ratib, "OsiriX: An open-source software for navigating in multidimensional DICOM images," *Journal of Digital Imaging*, vol. 17, no. 3, pp. 205–216, Sep. 2004. [Online]. Available: http://osirix-viewer.com/. doi:10.1007/s10278-004-1014-6

[32] E. Catmull and J. Clark, "Recursively generated B-spline surfaces on arbitrary topological meshes," *Computer-Aided Design*, vol. 10, no. 6, pp. 350–355, Nov. 1978, doi:10.1016/0010-4485(78)90110-0.

[33] Blender. [Online]. Available: http://blender.org/

[34] MeshLab. [Online]. Available: http://meshlab.sourceforge.net/

[35] S. Narayanan, E. Bresch, P. K. Ghosh, L. Goldstein, A. Katsamanis, Y. Kim, A. Lammert, M. Proctor, V. Ramanarayanan, and Y. Zhu, "A multimodal real-time MRI articulatory corpus for speech research," in *Interspeech*, Florence, Italy, Aug. 2011, pp. 837–840. [Online]. Available: http://www.isca-speech.org/archive/interspeech_2011/i11_0837.html

[36] A. Niebergall, S. Zhang, E. Kunay, G. Keydana, M. Job, M. Uecker, and J. Frahm, "Real-time MRI of speaking at a resolution of 33 ms: Undersampled radial FLASH with nonlinear inverse reconstruction," *Magnetic Resonance in Medicine*, vol. 69, no. 2, pp. 477–485, Feb. 2013, doi:10.1002/mrm.24276.

[37] Bullet physics library. [Online]. Available: http://bulletphysics.org/

[38] T. Kitamura, H. Takemoto, K. Honda, Y. Shimada, I. Fujimoto, Y. Syakudo, S. Masaki, K. Kuroda, N. Oku-Uchi, and M. Senda, "Difference in vocal tract shape between upright and supine postures: Observations by an open-type MRI scanner," *Acoustical Science and Technology*, vol. 26, no. 5, pp. 465–468, 2005, doi:10.1250/ast.26.465.