PROGRESS IN ANIMATION OF AN EMA-CONTROLLED TONGUE MODEL FOR ACOUSTIC-VISUAL SPEECH SYNTHESIS

Ingmar Steiner, Slim Ouni

LORIA Speech Group
Bat. C, 615 Rue du Jardin Botanique, 54600 Villers-lès-Nancy, France

Firstname.Lastname@loria.fr

Abstract: We present a technique for the animation of a 3D kinematic tongue model, one component of the talking head of an acoustic-visual (AV) speech synthesizer. The skeletal animation approach is adapted to make use of a deformable rig controlled by tongue motion capture data obtained with electromagnetic articulography (EMA), while the tongue surface is extracted from volumetric magnetic resonance imaging (MRI) data. Initial results are shown and future work outlined.

1 Introduction

As part of ongoing research in developing a fully data-driven acoustic-visual (AV) text-to-speech (TTS) synthesizer [16], we integrate a tongue model to increase visual intelligibility and naturalness. To extend the kinematic paradigm used for facial animation in the synthesizer to tongue animation, we adapt state-of-the-art techniques of animation with motion-capture data for use with electromagnetic articulography (EMA).

Our AV synthesizer[1] is based on a non-uniform unit-selection TTS system for French [4], concatenating bimodal units of acoustic and visual data, and extending the selection algorithm with visual target and join costs [13]. The result is an application whose graphical user interface (GUI) features a "talking head" (i.e. computer-generated face), which is animated synchronously with the synthesized acoustic output.

This synthesizer depends on a speech corpus acquired by tracking marker points painted onto the face of a human speaker, using a stereoscopic high-speed camera array, with simultaneously recorded audio. While the acoustic data is used for waveform concatenation in a conventional unit-selection paradigm, the visual data is post-processed to obtain a dense, animated 3D point cloud representing the speaker's face. The points are interpreted as the vertices of a mesh, which is then rendered as an animated surface to generate the face of the talking head using a standard *vertex animation* paradigm.

Due to the nature of the acquisition setup, no intra-oral articulatory motion data can be simultaneously captured. At the very least, any invasive instrumentation, such as EMA wires or transducer coils, would have a negative effect on the speaker's articulation and hence, the quality of the recorded audio; additional practical issues (e.g. coil detachment) would limit the length of the recording session, and by extension, the size of the speech corpus. As a consequence, the synthesizer's talking head currently features neither tongue nor teeth, which significantly decreases both the naturalness of its appearance and its visual intelligibility.

To address this shortcoming, we develop an independently animated 3D tongue and teeth model,

[1] `http://visac.loria.fr/`

which will be integrated into the talking head and eventually controlled by interfacing directly with the TTS synthesizer.

2 EMA-based tongue model animation

To maintain the data-driven paradigm of the AV synthesizer, the tongue model[2] consists of a geometric mesh rendered in the GUI along with (or rather, "behind") the face. Since the primary purpose of the tongue model is to improve the *visual* aspects of the synthesizer and it has no influence on the acoustics, there is no requirement for a complex tongue model to calculate the vocal tract transfer function, etc. Therefore, in contrast to previous work [e.g. 5, 6, 8, 10, 14, 17], most of which attempts to predict tongue shape and/or motion by simulating the dynamics in one form or another, we must merely generate realistic tongue *kinematics*, without having to model the anatomical structure of the human tongue or satisfy physical or biomechanical constraints.

This scenario allows us to make use of standard animation techniques using motion capture data. Specifically, we apply electromagnetic articulography (EMA) using a Carstens AG500[3] to obtain high-speed (200 Hz), 3D motion capture data of the tongue during speech [7].

While other modalities might be used to acquire the shape of the tongue while speaking, their respective drawbacks make them ill-suited to our needs. For example, ultrasound tongue imaging tends to require extensive processing to track the mid-sagittal tongue contour and does not usually capture the tongue tip, while real-time magnetic resonance imaging (MRI) has a very low temporal resolution, and is currently possible only in a single slice.[4]

2.1 Tongue motion capture

Conventional motion capture modalities (as widely used e.g. in the animation industry) normally employ a camera array to track optical markers attached to the face or body of a human actor, producing data in the form of a 3D point cloud sampled over time. For facial animation, these points (given sufficient density) can be directly used as vertices of a mesh representing the surface of the face; this is the vertex animation approach taken in the AV synthesizer (see above).

For articulated body animation, however, the 3D points are normally used as transformation targets for the rigid bones of a hierarchically structured (usually humanoid) skeleton model. Much like the strings controlling a marionette, the skeletal transformations are then applied to a virtual character by deforming its geometric mesh accordingly, a widely used technique known as *skeletal animation*.

Since current EMA technology allows the tracking of no more than 12 transducer coils (usually significantly fewer on the tongue), the resulting data is too sparse for vertex animation of the tongue surface. For this reason, we adopt a skeletal animation approach, but without enforcing a rigid structure, since the human tongue contains no bones and is extremely deformable. This issue is addressed below.

One advantage of EMA lies in the fact that the data produced is a set of 3D *vectors*, not points, as the AG500 tracks the orientation, as well as position, of each transducer coil. Thus, the rotational information supplements, and compensates to some degree for the sparseness of,

[2] For reasons of brevity, in the remainder of this paper, we will refer only to a *tongue* model, but it should be noted that that such a model can easily encompass upper and lower teeth in addition to the tongue.

[3] Carstens Medizinelektronik GmbH, `http://www.articulograph.de/`

[4] 3D cine-MRI of the vocal tract [15], while possible, is far from realistic for the compilation of a full speech corpus sufficient for TTS.

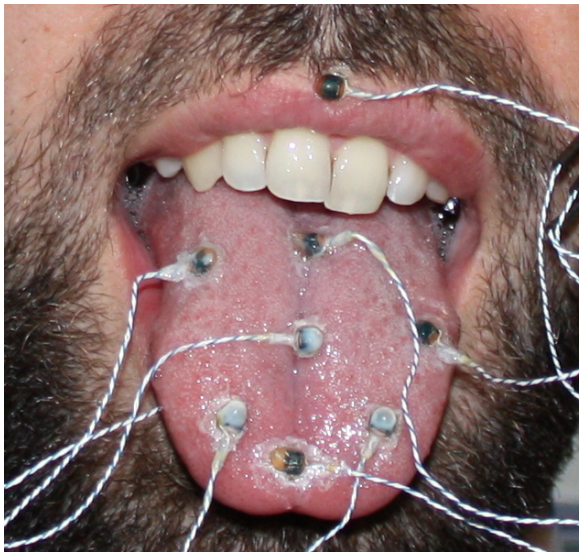


Figure 1 – Coil layout for EMA test corpus

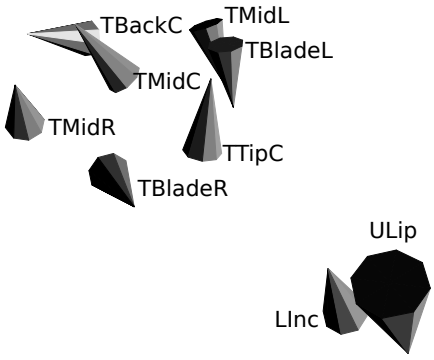


Figure 2 – Perspective view of EMA coils rendered as primitive cones to visualize their orientation

the positional data. Technically, this corresponds to motion capture approaches such as [11], although the geometry is of course quite different for the tongue than for a humanoid skeleton.

As a small EMA test corpus, we recorded one speaker using the AG500, with the following measurement coil layout: tongue tip center, tongue blade left/right, tongue mid center/left/right, tongue back center, lower incisor, upper lip (reference coils on bridge of nose and behind each ear). The exact arrangement can be seen in Figure 1. The speech material comprises sustained vowels in the set [i, y, u, e, ø, o, ə, a], repetitive CV syllables permuting these vowels with the consonants in the set [p, t, k, m, n, ŋ, f, θ, s, ʃ, ç, x, l, ɬ], as well as 10 normal sentences in German and English, respectively. A 3D palate trace was also obtained.

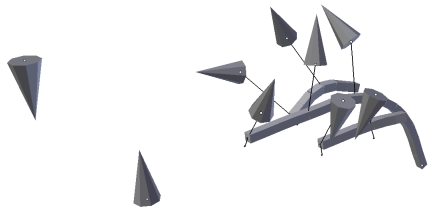
We imported the raw EMA data as keyframes into the animation component of a fully-featured, open-source, 3D modeling and animation software suite,[5] using a custom plugin. Unlike point cloud based motion capture data contained in industry standard formats such as C3D [12], this allows us to directly import the rotational data as well. As an example of the result, one frame is displayed in Figure 2. Within each frame of the animation, the EMA coil objects can provide the transformation targets for an arbitrary skeleton.

Once the motion capture data has been imported into the 3D software, it can be segmented into distinct actions for use and re-use in non-linear animation (NLA). This allows us to manipulate and concatenate any number of frame sequences as atomic actions, and to synthesize new animations from them, using e.g. the 3D software's NLA editor (which, for these purposes, is conceptually similar to a gestural score in articulatory phonology [3]).

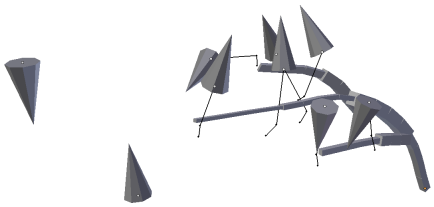
2.2 Tongue model animation

In order to use the tongue motion capture data to control a tongue model using skeletal animation, we design a simple skeleton as a rig for the tongue mesh. This rig consists of a central "spine", and two branches to allow (potentially asymmetric) lateral movement, such as grooving. Once again, it must be pointed out that this rig is unrelated to real tongue anatomy, although

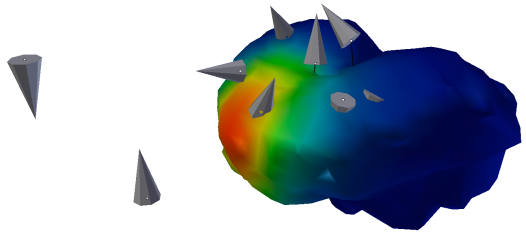
[5] Blender v2.5, `http://www.blender.org/`



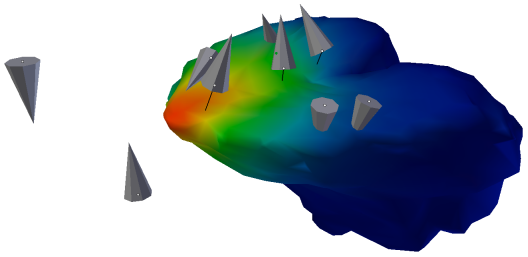
(a) Perspective view of EMA coils (rendered as cones) and deformable skeletal rig in bind pose ([a] vowel). Tongue tip at center, oriented towards left; upper lip and lower incisor coils are visible further left. Adaptation struts (IK targets, cf. text) are shown as thin rods connecting coils and B-bones



(b) Like 3a, but deformed according to [t] target pose



(c) Like 3a, but with the tongue mesh bound to the rig. Its surface is color-shaded with a heat map visualizing the influence of the tongue-tip B-bone on the mesh vertices (red=full; blue=none)



(d) Like 3c, but deformed according to [t] target pose

(e) Like 3c, but showing only the tongue surface mesh, with the tip at center, oriented left

(f) Like 3e, but deformed according to [t] target pose

Figure 3 – Tongue model in initial and final frame of one [at] cycle in the EMA test corpus

it could be argued that e.g. the spine corresponds roughly to the superior and inferior longitudinal muscles.

Of course, a skeleton of rigid bones is inadequate to mimic the flexibility of a real tongue. Our solution is to construct the rig using *deformable* bones, so-called bézier bones (B-bones), which can bend, twist, and stretch as required, governed by a set of constraining parameters.

The tongue model should be able to move independently of any specific EMA coil layout, since after all, the motion capture data represents *observations* of tongue movements based on hidden dynamics. For this reason, and to maintain as much modularity and flexibility as possible in the design, the animation rig is not directly connected to the EMA coils in the motion capture data. Instead, we introduce an *adaptation layer* in the form of "struts", each of which is connected to one coil, while the other end serves as a target for the rig's B-bones. These struts can be adapted to any given EMA coil layout or rig structure.

With the struts in place and constrained to the movements of the EMA coils, the rig can be animated by using *inverse kinematics (IK)* to determine the location, rotation, and deformation of each B-bone for any given frame. The IK are augmented by volume constraints, which inhibit potential "bloating" of the rig during B-bone stretching.

The final component for tongue model animation is a mesh that represents the tongue surface, which is rendered in the GUI and deformed according to the skeletal animation. While this tongue mesh could be an arbitrary geometric structure, we use an isosurface extracted from a volumetric scan in a MRI speech corpus (from a different speaker; voxel size $1.09\,\text{mm} \times 1.09\,\text{mm} \times 4\,\text{mm}$). The tongue in this scan was manually segmented using a graphics tablet and open-source medical imaging software.[6]

The resulting tongue mesh was manually registered to the EMA coil positions in a neutral *bind pose*. The skeletal rig was then embedded, and vertex groups in the tongue mesh assigned automatically to each B-bone. As the motion capture data animates the rig using IK, the tongue mesh is deformed accordingly, approximately following the EMA coils. Figure 3 displays the initial and final frame in one cycle of repetitive [ta] articulation in the EMA test corpus. In an informal evaluation, our technique appears to produce satisfactory results, and encourages us to pursue and refine this approach to tongue model animation.

3 Discussion and Outlook

We have presented a technique to animate a kinematic tongue model, based on volumetric vocal tract MRI data, using skeletal animation with a flexible rig, controlled by motion capture data acquired with EMA, and implemented with off-the-shelf, open-source software. While this approach appears promising, it is still under development, and there are various issues which must be addressed before the tongue model can be integrated into our AV TTS synthesizer as intended.

- Upper and lower teeth can be added to the model using the same data and animation technique, albeit with a conventional, rigid skeleton. These will then be rendered in the synthesizer's GUI along with the face and tongue.

- The tongue mesh used here is quite rough, and registration with the EMA data does not produce the best fit, owing to differences between the speakers' vocal tract geometries and articulatory target positions, quite possibly exacerbated by the effects of supine posture

[6] OsiriX v3.9, `http://www.osirix-viewer.com/`

during MRI scanning [e.g. 9]. A more suitable mesh might be obtained by scanning the tongue of the same speaker used for the EMA motion capture data, at a higher resolution.

- Registration of the tongue mesh into the 3D space of the tongue model should be performed in a partially or fully automatic way, using landmarks available in both MRI and EMA modalities [cf. 1], such as the 3D palate trace and/or high-contrast markers at the positions of the reference coils.

- The reliability of EMA position and orientation data is sometimes unpredictable. This could be due to the algorithms used to process the raw amplitude data, faulty hardware, interference (even within the coil layout itself), or any combination of such factors. However, since any such errors are immediately visible in the animation of the tongue model by introducing implausible deformations, we are working on methods both to clean the EMA data itself, and to make the tongue model less susceptible to such outlier trajectory segments.

- To evaluate the performance of the animation technique, factors such as skin deformation and distance of EMA coils from the tongue model surface should be monitored. The structure of the skeletal rig can be independently refined, optimizing its ability to generate realistic tongue poses. Its embedding into the tongue mesh should preferably be performed using a robust automatic method [e.g. 2].

- The 3D palate trace can be used to add a palate surface mesh to the tongue model. For both the palate and the teeth, the model could also be augmented with automatic collision detection by accessing the 3D software's integrated physics engine.[7]

For an interactive application such as the AV synthesizer GUI, it is impractical to incur the performance overhead of an elaborate 3D rendering engine, especially when a non-trivial processing load is required for the bimodal unit-selection. Instead, we anticipate integrating the tongue model into the talking head using a more lightweight, real-time capable 3D game engine, which may even offload the visual computation to dedicated graphics hardware. The advantage of using keyframe-based, NLA actions is that they can be ported into such engines as animated 3D models, using common interchange formats.[8] Although the skeletal rig could be accessed and manipulated directly, this "pre-packaging" of animation actions also avoids the complexity, or perhaps even unavailability, of advanced features such as B-bones or IK in those engines.

The final integration challenge is to interface the tongue model directly with the TTS system to synthesize the correct animation actions with appropriate timings. This task might be accomplished using a diphone synthesis style approach, or even action unit-selection, and will be addressed in the near future.

Acknowledgments

We owe our thanks to Sébastien Demange for assistance during the recording of the EMA test corpus, and to Korin Richmond for providing the means to record the MRI data used here.

[7] Bullet physics library, `http://www.bulletphysics.com/`

[8] For instance, COLLADA (`http://www.collada.org/`) or OGRE (`http://www.ogre3d.org/`)

References

[1] ARON, M., A. TOUTIOS, M.-O. BERGER, E. KERRIEN, B. WROBEL-DAUTCOURT and Y. LAPRIE: *Registration of multimodal data for estimating the parameters of an articulatory model*. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4489–4492, Taipei, Taiwan, April 2009. IEEE.

[2] BARAN, I. and J. POPOVIĆ: *Automatic rigging and animation of 3D characters*. In *Proc. 34th International Conference and Exhibition on Computer Graphics and Interactive Techniques (SIGGRAPH)*, San Diego, CA, USA, August 2007. ACM.

[3] BROWMAN, C. and L. GOLDSTEIN: *Articulatory Phonology: An overview*. Phonetica, 49(3-4):155–180, 1992.

[4] COLOTTE, V. and R. BEAUFORT: *Linguistic features weighting for a text-to-speech system without prosody model*. In *Proc. Interspeech*, pp. 2549–2552, Lisbon, Portugal, September 2005. ISCA.

[5] ENGWALL, O.: *Combining MRI, EMA & EPG measurements in a three-dimensional tongue model*. Speech Communication, 41(2-3):303–329, October 2003.

[6] GERARD, J.-M., P. PERRIER and Y. PAYAN: *3D biomechanical tongue modeling to study speech production*. In HARRINGTON, J. and M. TABAIN (eds.): *Speech Production: Models, Phonetic Processes, and Techniques*, chap. 10, pp. 149–164. Psychology Press, New York, NY, May 2006.

[7] KABURAGI, T., K. WAKAMIYA and M. HONDA: *Three-dimensional electromagnetic articulography: A measurement principle*. Journal of the Acoustical Society of America, 118(1):428–443, July 2005.

[8] KING, S. A. and R. E. PARENT: *Creating speech-synchronized animation*. IEEE Transactions on Visualization and Computer Graphics, 11(3):341–352, May/June 2005.

[9] KITAMURA, T., 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, 26(5):465–468, 2005.

[10] LU, X. B., W. THORPE, K. FOSTER and P. HUNTER: *From experiments to articulatory motion – a three dimensional talking head model*. In *Proc. Interspeech*, pp. 64–67, Brighton, UK, September 2009. ISCA.

[11] MOLET, T., R. BOULIC and D. THALMANN: *Human motion capture driven by orientation measurements*. Presence: Teleoperators and Virtual Environments, 8(2):187–203, April 1999.

[12] MOTION LAB SYSTEMS: *The C3D File Format User Guide*. Baton Rouge, LA, USA, 2008.

[13] MUSTI, U., V. COLOTTE, A. TOUTIOS and S. OUNI: *Introducing visual target cost within an acoustic-visual unit-selection speech synthesizer*. In *Proc. 10th International Conference on Auditory-Visual Speech Processing (AVSP)*, Volterra, Italy, September 2011. ISCA.

[14] PELACHAUD, C., C. VAN OVERVELD and C. SEAH: *Modeling and animating the human tongue during speech production*. In *Proc. Computer Animation*, pp. 40–49, Geneva, Switzerland, May 1994. IEEE.

[15] TAKEMOTO, H., K. HONDA, S. MASAKI, Y. SHIMADA and I. FUJIMOTO: *Measurement of temporal changes in vocal tract area function from 3D cine-MRI data*. Journal of the Acoustical Society of America, 119(2):1037–1049, February 2006.

[16] TOUTIOS, A., U. MUSTI, S. OUNI, V. COLOTTE, B. WROBEL-DAUTCOURT and M.-O. BERGER: *Towards a true acoustic-visual speech synthesis*. In *Proc. 9th International Conference on Auditory-Visual Speech Processing (AVSP)*, pp. POS1–8, Hakone, Kanagawa, Japan, September 2010. ISCA.

[17] VOGT, F., J. E. LLOYD, S. BUCHAILLARD, P. PERRIER, M. CHABANAS, Y. PAYAN and S. S. FELS: *An efficient biomechanical tongue model for speech research*. In *Proc. 7th International Seminar on Speech Production (ISSP)*, pp. 51–58, Ubatuba, Brazil, December 2006.