Towards an articulatory tongue model using 3D EMA

Ingmar Steiner[1,2], **Slim Ouni**[1,3]

[1]LORIA Speech Group

[2]INRIA Grand Est

[3]Université Nancy 2
Bat. C, 615 Rue du Jardin Botanique, 54600 Villers-lès-Nancy, France

Firstname.Lastname@loria.fr

Abstract. *Within the framework of an acoustic-visual (AV) speech synthesizer, we describe a preliminary tongue model that is both simple and flexible, and which is controlled by 3D electromagnetic articulography (EMA) data through an animation interface, providing realistic tongue movements for improved visual intelligibility. Data from a pilot study is discussed and deemed encouraging, and the integration of the tongue model into the AV synthesizer is outlined.*

1. Introduction

It is well known that the human tongue plays a significant role in speech production. It participates in several kinds of sounds, and its position is critical for several, if not most, phonemes. Even though it is only partially visible from outside the mouth, the tongue helps to improve the intelligibility of audiovisual speech. This is especially important for lip reading since it gives useful information for dental and liquid sounds, for example. Within the framework of developing a talking head, adding a tongue model increases drastically the overall intelligibility of the visual articulation, compared to talking heads displaying only the face and lips.

The tongue is a complex organ that is highly flexible, expandable, compressible, can be curved and displays a fine degree of articulation (e.g. Abd-el-Malek, 1939; Bole and Lessler, 1966; Carpentier and Pajoni, 1989). The tongue and its muscles are laterally symmetrical: a median septum divides the organ into two halves. The dynamic aspects of the tongue articulation (mainly coarticulation) are important as this has an influence on the intelligibility of the overall gestures.

In previous work, several approaches have been taken to model the shape, as well as the dynamic properties, of the tongue during speech, with varying complexity (e.g. Pelachaud et al., 1994; King and Parent, 2005; Gerard et al., 2006; Lu et al., 2009), and while many use vocal tract MRI to obtain static tongue shapes, the dynamics of running speech are more successfully observed using modalities such as electromagnetic articulography (EMA). Most specifically, Engwall (2003) presents a MRI-adapted, 3D parametric tongue model, which can be externally controlled by EMA data, but this data is only two-dimensional in nature, and the mapping from the EMA coils to model control parameters is somewhat simplified.

Within the scope of a project developing an acoustic-visual (AV) speech synthesizer,[1] our aim is to implement and integrate a tongue model into the synthesizer's animated talking head that will improve the intelligibility and naturalness of the resulting AV speech output. In particular, the movements of the tongue model should appear as natural as possible.

Modeling a realistic tongue with high precision in articulation presents a significant challenge. In developing the AV synthesizer, we mainly focus on *visual* intelligibility. In this context, the tongue is partially visible and thus we consider that the dynamics are more important than finely modeling the shape or position of the tongue. For these reasons, the objective of this work is to build a simple dynamic tongue model.

The AV synthesizer concatenates units selected from a multimodal corpus of facial motion-capture data and simultaneous acoustic recordings (Toutios et al., 2010). The motion-capture was performed with a stereoscopic high-speed camera array tracking 252 facial marker points at a rate approaching 200 Hz. The acoustic data is processed and used directly for speech synthesis by waveform concatenation in a unit selection paradigm.

Apart from the facial motion-capture, no further articulatory data is available for the AV corpus, because the instrumentation of modalities such as EMA would have interfered with the optical acquisition. While simultaneous optical and articulatory tracking has been demonstrated in the past (e.g. Yehia et al., 1998; Engwall and Beskow, 2003; Kroos, 2008; Lu et al., 2009), invasive components such as wires and transducer coils would have significantly affected the quality of the acoustic recordings, and the duration of the recording session (and thus, the size of the resulting multimodal corpus) would have been limited by EMA-inherent practicalities (e.g. coil detachment).

Nevertheless, with the AV synthesizer driven entirely by data captured from a real human speaker, we explore the possibility of animating the tongue through articulatory data in an analogous manner. Of course, for this to be possible, an auxiliary corpus of such data must be available, and multimodal synthesis from both corpora must be synchronized and merged appropriately. The former precondition is the subject of this paper, while the integration of facial/acoustic and tongue movement synthesis will be addressed in a later stage of the project.

The key difference between the facial motion-capture data in the primary AV corpus and tongue surface data in the auxiliary articulatory corpus is one of granularity. The 252 marker points represent the vertices of a mesh sufficiently fine to serve as the basis for the surface of the synthetic face. The movements of the tongue surface, on the other hand, cannot be tracked using such a fine sampling, since the technical limitations of EMA restrict the number of transducer coils. Therefore, EMA offers high temporal resolution, but at best only a *sparse* representation of the tongue surface.[2] Conversely, the five-dimensional measurement data of the Carstens AG500 EMA system provides not only the position of each coil in space, but also its orientation (Kaburagi et al., 2005).

In this paper, we present how the 3D positions and the 2 rotation angles of the EMA data can be used to provide information for tongue animation. We start from a very

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

[2] Some previous studies (Kaburagi and Honda, 1994; Qin and Carreira-Perpiñán, 2010) have explored the possibility of reconstructing the tongue contour from EMA data, but only in the mid-sagittal plane.

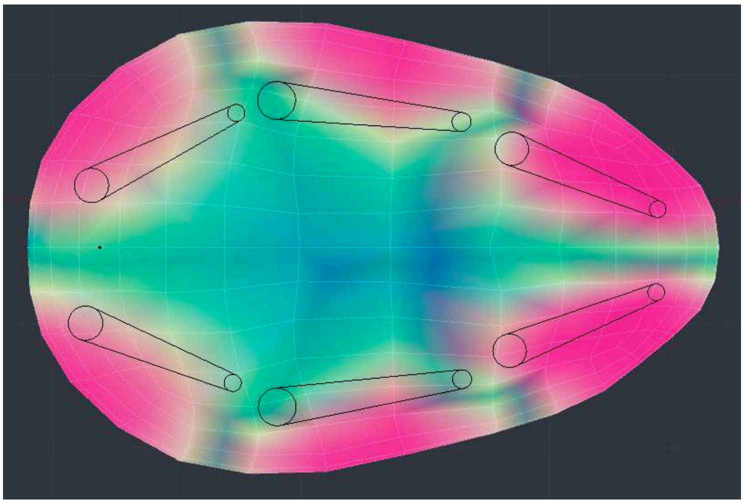

Figure 1. Top view of armature (six components) and an automatically assigned vertex weight map, with the colors indicating the degree of influence of each component on the vertices; the tongue mesh is visible as a wireframe (tip on right).

simple geometric tongue model. Then, the articulatory data, acquired using the AG500, is used to control the tongue shape.

2. Tongue model and Animation

To integrate a tongue model into the AV synthesizer, using the same approach for tongue and facial animation, articulatory point tracking data at a high temporal resolution is required. However, unlike the facial motion-capture data, which provides 3D movement data for hundreds of points on the speaker's face, enough to use these points directly as the vertices of the facial mesh in the talking head, only a very small number of transducer coils (by comparison), and hence, fleshpoints on the tongue can be tracked using EMA. The resulting data yields an insufficient number of vertices to be used for the tongue surface in the same way as the AV motion-capture data for the face.

To address this issue, we reduce the dimensionality of the tongue control in such a way that the tongue mesh is not wholly defined by, but only controlled by, the EMA data. This corresponds to the fact that the human speaker's tongue is not controlled by the EMA coils, but that the coils represent observations of fleshpoints on the tongue surface, and move in synchrony with them (by virtue of being physically attached).

An advantage of 3D EMA which counters some of the drawbacks of data sparsity, lies in the fact that along with the spatial positions of the transducer coils, their orientation in spherical coordinates is measured as well. This additional information is well-suited for transformational projection onto geometric structures larger than the coils themselves.

The tongue mesh itself is defined by a set of vertices arranged in a shape that happens to resemble that of a human tongue, but could in fact just as well be a sphere, cylinder, or similar construct. The critical fact is that it *moves* in unison with the EMA data controlling it. Each EMA coil can be associated with a subset of the tongue mesh

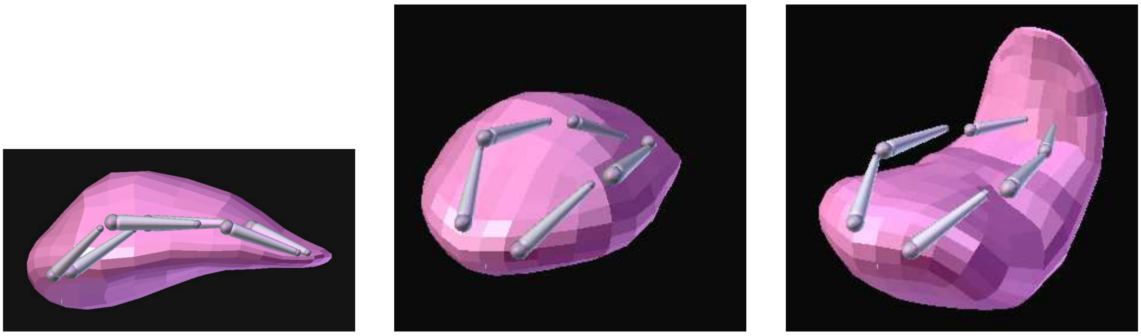

Figure 2. Side view (slightly oblique) of tongue mesh (tip on right) and armature, superimposed as six stick-like shapes, and two deformations of the mesh (tongue back towards the viewer) obtained by manipulating the armature. The center and right panels show a bunched and retroflex configuration of the tongue model, respectively.

vertices, and this association can furthermore be weighted either individually or by vertex groups. Hence, the movement and rotation of each EMA coil can be transferred to the vertices, thereby animating the tongue mesh. This can be described as a form of the *skeletal animation* technique.

However, both the position on the surface of the tongue and the small, point-like *shape*, of the EMA coils limit the suitability of the coils for direct use in skeletal animation. For this reason, we allow the skeletal animation to be controlled by the EMA coils only indirectly, by means of a weighted mapping to a pseudo-skeleton, or articulated *armature*, rigged inside the tongue mesh in a neutral *bind pose*. In line with conventional skeletal animation, the armature is in turn mapped to vertex groups on the tongue mesh by a weight map which can be defined automatically from the envelope of the armature's shape and manually adjusted if required, and manipulating the armature's components deforms the surrounding mesh accordingly. Figure 1 shows a simple armature and its automatic vertex weight map, where the influence of the armature's components on the mesh is visualized on a color scale, ranging from blue (low) to red (high).

The resulting tongue animation preserves the high temporal resolution of the controlling EMA data, and is both simple and flexible, as shown in Figure 2. Moreover, the details of the animation can be fine-tuned through the armature-coil mapping, as well as through the armature structure itself, and the vertex groups associated with it. In addition, the modular design allows the simple structure of the preliminary tongue model to be refined and improved at a later stage by replacing the mesh and rebinding the armature.

A distinct advantage of this approach to tongue modeling for AV synthesis lies in the fact that it employs well-developed and state-of-the-art concepts and techniques, and can be prototyped with relative ease using off-the-shelf 3D modeling software and engines, even addressing more advanced issues such as collision detection or soft body dynamics, by accessing the corresponding physics subsystems.

2.1. Pilot study

As a pilot study and to provide preliminary data during the development of the tongue model, a small corpus of 3D EMA data was recorded using the Carstens AG500 Artic-

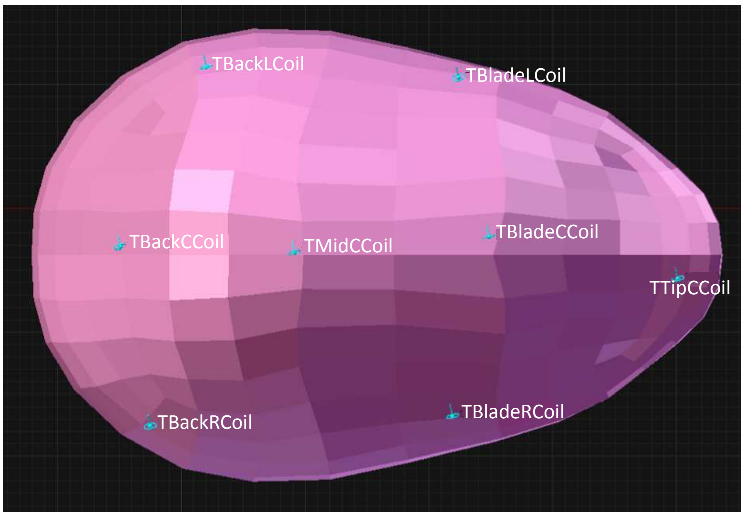

Figure 3. Top view of tongue model (tip on right) and layout of EMA coils on the tongue in pilot study. Coils, shown in aqua color, with a short line indicating the orientation, along the center (left to right): TBackC, TMidC, TBladeC, TTipC, and TBackL, TBladeL (top) and TBackR, TBladeR (bottom).

ulograph. Since our focus was on the tongue surface, there was no need to record lip movements; instead, a total of eight coils were attached to the speaker's tongue: four coils in the mid-sagittal plane (tongue tip, blade, mid, and back center), as well as two on each side (tongue blade left/right, and tongue back left/right); the resulting layout is shown with the tongue model in Figure 3. In addition, the jaw aperture was captured by a coil on the lower incisors. Reference coils for the normalization with respect to head movement were placed on the upper incisors, and behind both ears. However, using the alternative normalization approach proposed by Kroos (2009), it might be possible to increase the number of available measurement coils on the tongue even further by using fewer reference coils.

The unusually large number of tongue coils and their corresponding wires presented a practical challenge, and we did not manage to attach all of the coils in the optimal orientation. Ideally, for coils arranged along the mid-sagittal contour of the tongue, the coil axes should lie in the mid-sagittal plane, while for the lateral coils, a transverse axis orientation tends to provide a more robust indication of tongue grooving. In addition, we are currently uncertain whether the close proximity of coils on the tongue may have interfered with the measurement accuracy for some of the EMA channels (see below).

The recorded material includes CVC sequences, several utterances from a French prompt list, a 3D palate trace (using the LInc coil, attached to a pen), and one sweep during which the speaker explored various unusual tongue configurations without speaking.

This EMA data was visualized in a 3D environment, with the tongue model animated through the interface described above. Preliminary assessment seems to indicate that the tongue model can indeed be animated in a satisfactory way using the design described here, but a number of issues remain to be addressed. For instance, the optimal smoothing of the noise in the raw EMA trajectories must be determined to remove visible jitter without negatively affecting the animation. Another issue is the reliability of the EMA data; a few coils in the pilot data exhibit intermittent out-of-bounds positions.

Such unrealistic movements correlate with RMS spikes for the corresponding channels, but some dimensions seem to be more sensitive than others. The source of the problem might be related to the coil hardware itself, or the coil layout (in particular, their orientation and close proximity), or perhaps the method used to extract the positions from the raw data. While it may be possible to debug and reduce such errors, it is more than likely that they cannot be fully eliminated, and so the tongue model's control interface must be extended to robustly and gracefully handle these invalid positions.

2.2. Integration

While the tongue model is yet to be integrated with the AV synthesizer described in Section 1, several approaches are conceivable to generate the EMA trajectories required to control the tongue model for unseen utterances, i.e. within the final text-to-speech (TTS) application. All of them make use of a planned corpus, containing EMA recordings for a short prompt list providing at least minimal phonetic coverage.

By applying acoustic-to-articulatory inversion techniques, it may be possible to enrich the full AV corpus with reconstructed articulatory data *offline*, which can be used at synthesis time for tongue model animation in a way analogous to EMA data. However, it remains to be explored whether the additional dimensions required can be recovered in a satisfactory way by inversion techniques (most of which were developed using 2D EMA). If such an approach proved feasible, then the integration of the tongue model into the multimodal unit selection synthesizer would be straightforward, as corresponding articulatory trajectories would be bundled with every selected unit and could be used to animate the tongue along with the face for the corresponding segments of synthesized speech.

An alternative approach for the generation of articulatory control data lies in *trajectory synthesis* to generate the required control data at synthesis time, either using statistical models (e.g. Ling et al., 2010) or by direct concatenation (e.g. Engwall, 2002) in a separate, but parallel unit-selection thread fed by the auxiliary corpus. In the latter case, the units would not necessarily have to match those selected from the primary AV corpus for speech and facial animation, but could be handled in a syllable frame or as kinematic n-phones (Okadome and Honda, 2001).

3. Conclusion

We have described a tongue model designed for integration into an AV speech synthesizer, whose primary goal is to increase the visual intelligibility of a talking head. The model is simple but flexible, composed of a 3D mesh controlled by 3D EMA data through a skeletal animation interface. Initial tests with a small corpus of pilot data show encouraging results.

Future work includes refining the interface to robustly handle noisy data with intermittent measurement errors, recording and processing a larger, phonetically balanced EMA corpus, and integrating the tongue model into the AV synthesizer. An open question is whether the animation of the face and tongue, which so far are completely separate, could be partially combined, perhaps with one or both informing the other (e.g. Yehia et al., 1998; Engwall and Beskow, 2003). Finally, an evaluation study is planned to assess the performance of the tongue model.

References

Abd-el-Malek, S. Observations on the morphology of the human tongue. *Journal of Anatomy*, 73(pt.2):201–210, January 1939. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1252504/`.

Bole, C. T. and Lessler, M. A. Electromyography of the genioglossus muscles in man. *Journal of Applied Physiology*, 21(6):1695–1698, November 1966. URL `http://jap.physiology.org/content/21/6/1695.extract`.

Carpentier, P. and Pajoni, D. La langue: un ensemble musculaire complexe. *Revue d'Orthopédie Dento-Faciale*, 23(1):19–28, March 1989. doi:10.1051/odf/1989006.

Engwall, O. Evaluation of a system for concatenative articulatory visual speech synthesis. In *Proc. 7th International Conference on Spoken Language Processing*, pages 665–668, Denver, Colorado, USA, September 2002. ISCA. URL `http://www.isca-speech.org/archive/icslp_2002/i02_0665.html`.

Engwall, O. Combining MRI, EMA & EPG measurements in a three-dimensional tongue model. *Speech Communication*, 41(2-3):303–329, October 2003. doi:10.1016/S0167-6393(02)00132-2.

Engwall, O. and Beskow, J. Resynthesis of 3D tongue movements from facial data. In *Proc. Eurospeech*, pages 2261–2264, Geneva, Switzerland, September 2003. ISCA. URL `http://www.isca-speech.org/archive/eurospeech_2003/e03_2261.html`.

Gerard, J.-M., Perrier, P., and Payan, Y. 3D biomechanical tongue modeling to study speech production. In Harrington, J. and Tabain, M., editors, *Speech Production: Models, Phonetic Processes, and Techniques*, chapter 10, pages 149–164. Psychology Press, New York, NY, May 2006. ISBN 978-1-84169-437-5.

Kaburagi, T. and Honda, M. Determination of sagittal tongue shape from the positions of points on the tongue surface. *Journal of the Acoustical Society of America*, 96(3): 1356–1366, September 1994. doi:10.1121/1.410280.

Kaburagi, T., Wakamiya, K., and Honda, M. Three-dimensional electromagnetic articulography: A measurement principle. *Journal of the Acoustical Society of America*, 118 (1):428–443, July 2005. doi:10.1121/1.1928707.

King, S. A. and Parent, R. E. Creating speech-synchronized animation. *IEEE Transactions on Visualization and Computer Graphics*, 11(3):341–352, May/June 2005. doi:10.1109/TVCG.2005.43.

Kroos, C. Measurement accuracy in 3D electromagnetic articulography (Carstens AG500). In *Proc. 8th International Seminar on Speech Production*, pages 61–64, Strasbourg, France, December 2008. LORIA. URL `http://issp2008.loria.fr/Proceedings/PDF/issp2008-9.pdf`.

Kroos, C. Using sensor orientation information for computational head stabilisation in 3D electromagnetic articulography (EMA). In *Proc. Interspeech*, pages 776–779, Brighton, UK, September 2009. ISCA. URL `http://www.isca-speech.org/archive/interspeech_2009/i09_0776.html`.

Ling, Z.-H., Richmond, K., and Yamagishi, J. HMM-based text-to-articulatory-movement prediction and analysis of critical articulators. In *Proc. Interspeech*, pages 2194–2197, Chiba, Japan, September 2010. ISCA. URL `http://www.isca-speech.org/archive/interspeech_2010/i10_2194.html`.

Lu, X. B., Thorpe, W., Foster, K., and Hunter, P. From experiments to articulatory motion – a three dimensional talking head model. In *Proc. Interspeech*, pages 64–67, Brighton, UK, September 2009. ISCA. URL `http://www.isca-speech.org/archive/interspeech_2009/i09_0064.html`.

Okadome, T. and Honda, M. Generation of articulatory movements by using a kinematic triphone model. *Journal of the Acoustical Society of America*, 110(1):453–463, 2001. doi:10.1121/1.1377633.

Pelachaud, C., van Overveld, C., and Seah, C. Modeling and animating the human tongue during speech production. In *Proc. Computer Animation*, pages 40–49, Geneva, Switzerland, May 1994. IEEE. doi:10.1109/CA.1994.324008.

Qin, C. and Carreira-Perpiñán, M. Á. Reconstructing the full tongue contour from EMA/X-ray microbeam. In *Proc. International Conference on Acoustics, Speech & Signal Processing*, pages 4190–4193, Dallas, TX, USA, March 2010. IEEE. doi:10.1109/ICASSP.2010.5495712.

Toutios, A., Musti, U., Ouni, S., Colotte, V., Wrobel-Dautcourt, B., and Berger, M.-O. Towards a true acoustic-visual speech synthesis. In *Proc. 9th International Conference on Auditory-Visual Speech Processing*, pages POS1–8, Hakone, Kanagawa, Japan, September 2010.

Yehia, H., Rubin, P., and Vatikiotis-Bateson, E. Quantitative association of vocal-tract and facial behavior. *Speech Communication*, 26(1-2):23–43, October 1998. doi:10.1016/S0167-6393(98)00048-X.