From Tongue Movement Data to Muscle Activation – A Preliminary Study of Artisynth's Inverse Modelling

Saeed Dabbaghchian
KTH
Lindstedtsvägen 24
SE-100 44, Stockholm, Sweden
saeedd@kth.se

Isak Nilsson
KTH
Lindstedtsvägen 24
SE-100 44, Stockholm, Sweden
isakni@kth.se

Olov Engwall
KTH
Lindstedtsvägen 24
SE-100 44, Stockholm, Sweden
engwall@kth.se

Abstract

Finding the muscle activations during speech production is an important part of developing a comprehensive biomechanical model of speech production. Although there are some direct ways, like Electromyography, for measuring muscle activations, these methods usually are highly invasive and sometimes not reliable. They are more over impossible to use for all muscles. In this study we therefore explore an indirect way to estimate tongue muscle activations during speech production by combining Electromagnetic Articulography (EMA) measurements of tongue movements and the inverse modeling in Artisynth. With EMA we measure the time-changing 3D positions of four sensors attached to the tongue surface for a Swedish female subject producing vowel-vowel and vowel-consonant-vowel (VCV) sequences. The measured sensor positions are used as target points for corresponding virtual sensors introduced in the tongue model of Artisynth's inverse modelling framework, which computes one possible combination of muscle activations that results in the observed sequence of tongue articulations. We present resynthesized tongue movements in the Artisynth model and verify the results by comparing the calculated muscle activations with literature.

Keywords: speech, tongue, muscle activation, electromagnetic articulography, biomechanics.

1. Introduction

Since speech is the most important tool for human communication, understanding the nature of the voice and its production mechanism has triggered research for centuries. A model of the human voice apparatus has numerous applications in different domains, e.g. helping clinicians to realize how voice problems arise, and how they should be diagnosed and treated. If the model is able to simulate generation of natural sounding voices, it could further both increase our knowledge of speech production and lead to methods to synthesize speech more flexibly. This is the aim of the on-going European project EUNISON (eunison.eu), which combines biomechanical modeling of the voice organs with fluid mechanics simulations of the wave propagation in the vocal tract.

The tongue has a particularly important role in speech production, since its position and shape determine the acoustic output for a large number of phonemes. Since it is hidden from normal view, various methods for measurement and visualization have been employed over the decades, including X-rays, Magnetic Resonance Imaging (MRI), Electromagnetic Articulography (EMA) and Electropalatography. The relatively rapid tongue movements are however a challenge for the measurement techniques. As MRI acquisition times have been decreased in recent years, it has been possible to acquire sequences of 2D images in the mid-sagittal plane. As an alternative, EMA is a simpler and faster method for recording tongue movements by point-wise tracking of sensors attached to the tongue surface [1][2][3]. As the points on the tongue cannot move independently because they are connected through the tissues and muscles, it is possible to recreate the entire tongue deformation using only 3-4 sensor [4][5].

A biomechanical model of tongue, which simulates the muscle contractions, can be used to find the entire tongue deformation during the speech production based on flesh-point measures. Gérard et al. [6] developed a biomechanical tongue model that simulates the muscle activation effect on the tongue shape by using the finite element method (FEM). Fang and et al. [7] developed a 3D tongue model to study the muscle activation and speech motor control. In order find muscle activation from measures of the tongue movements, an inverse problem must be solved, i.e. finding a combination of muscle activations that generates the observed tongue shapes. Stavness et al. developed an inverse tongue model [8] in which some points on the tongue are defined as targets. By assuming that the trajectories of these target points are known, the muscle activations are computed so that the tongue model follows the trajectories of the target points.

PMHA-14 Aug 22-23, 2014, Vancouver, BC, CA

After computing the muscle activations, the biomechanical model can predict the entire tongue deformation.

In this paper, we utilize the inverse tongue model in Artisynth as a biomechanical model and try to find the muscle activations and the entire tongue deformation consequently by using EMA data as trajectory for the target points. In the rest of this paper, section 2 describes the EMA data, section 3 the Artisynth inverse tongue model and section 4 the preprocessing of EMA data. Simulation results are presented in section 5 and discussed in section 6.

2. EMA data

We use EMA data of a native female speaker of Swedish producing phoneme sequences of vowel-vowel and vowel-consonant-vowel combinations, such as /ɑ:i:/, /ɑ:u:/, /i:u:/, /as:a/ and /ak:a/.

A Carstens AG500 articulograph [9] with twelve sensors is used to record the 3D position and orientation of each of the ears, the corners of the mouth, the upper and lower lips, the nose, the lower jaw and four sensors on the tongue (c.f. Figure 1). The four sensors on the tongue are placed on the surface in the mid-sagittal plane: one near the tongue tip, another at the back of the tongue and two others at the tongue body evenly spaced between the tip and back sensors. The sensors by the ears are used as references for compensating for the head movement. Although 3D positions and rotations for all twelve sensors are available, we use only the 3D positions of the sensors on the tongue, as input for the inverse tongue model, in this study.

The audio files of the utterances were synchronously recorded with sensor positions.

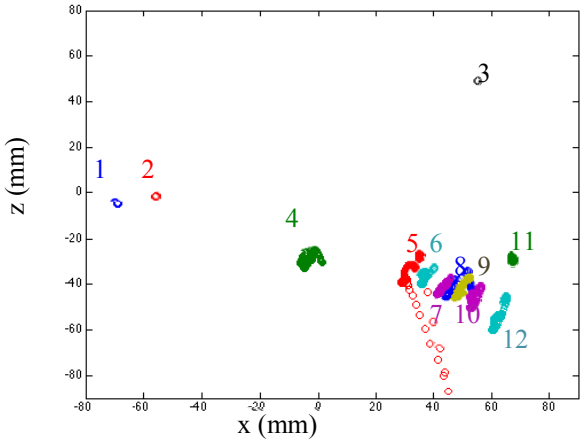

Figure 1. Midsaggital (x-z) view of the EMA data during /ak:a/ for the 12 sensor: (1-2) ears, (3) nose, (4, 5, 6, 8) tongue, (9) jaw, (7 and 10) lip corners, (11) upper lip and (12) lower lip.

3. Artisynth Inverse Tongue Model

Artisynth is a 3D biomechanical simulation environment, which allows the physical simulations of anatomical structures [10]. In this paper, we utilize the 3D finite element tongue model developed by Stavenes et al. [8]. In this model, eleven muscles control the tongue deformation, namely genioglossus (GG), styloglossus (SG), geniohyoid (GH), mylohyoid (MH), hyoglossus (HG), verticalis (V), transversus (T), inferior longitudinal (IL) and superior longitudinal (SL). The genioglossus was further divided into anterior (GGa), middle (GGm) and posterior (GGp). The placement and fibre directions of these eleven muscles are illustrated in Figure 2. In the tongue model the user can on the one hand change the muscle activations and observe the tongue deformation or movement. On the other, it is also possible to assign the trajectories of a number of markers on the tongue surface and let the model follow these trajectories optimally, by combining the activation of the different muscles so that the difference between the marker and target positions are minimized.

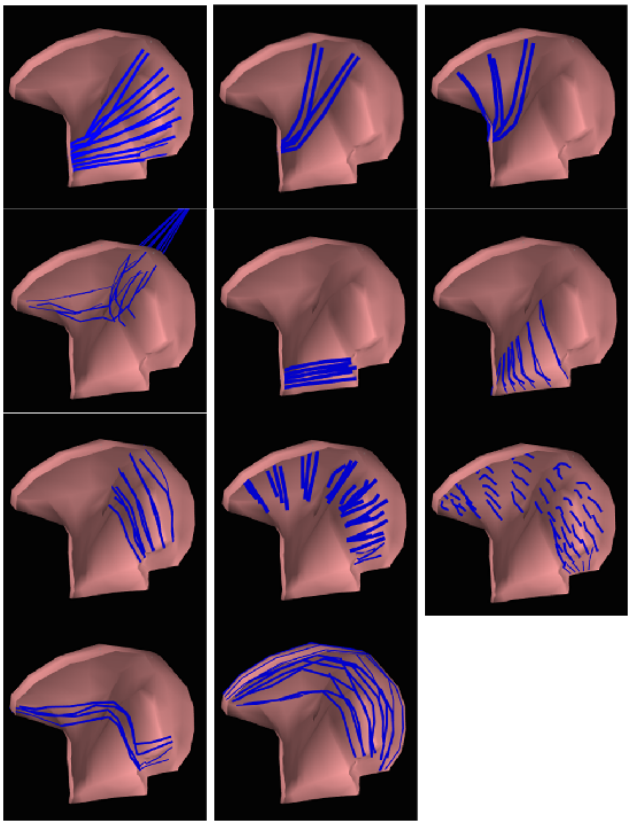

Figure 2. Placement and fibre directions in the tongue model of the muscles (in order from top left) genioglossus posterior (GGp), genioglossus middle (GGm), genioglossus anterior (GGa), styloglossus (SG), geniohyoid (GH), mylohyoid (MH), hyoglossus (HG), verticalis (V), transversus (TRANS), inferior longitudinal (IL) and superior longitudinal (SL).

4. Data Preprocessing

Since EMA measurements are noisy and have spurious measurement errors (see for example sensor 5 in Figure 1), a moving average filter is used to smooth the sensors movement. This filter averages the five previous samples in each time step.

Furthermore, we need to consider that the EMA data has been collected for one subject and the tongue model has been created from another subject. Therefore, another preprocessing of EMA data is required to compensate for the differences between the tongue sizes and shapes of the two subjects. To do this, the average of the EMA data during the silent period of the corresponding audio file is computed, assuming that these averages indicate the sensor positions when the tongue is in rest position. Then, these positions are mapped onto the mid-sagittal plane of the tongue model in Artisynth, so that the experimental set-up in the EMA data acquisition is replicated in the tongue model. Each EMA sensor is represented by a virtual sensor in the tongue model, with distances between sensors preserved (c.f. Figure 3). The computation of the positioning for each sensor is not automatic and needs user interaction. This is done only one time for each sensor; and once it is computed, the EMA data are updated to the new coordination system.

The transformed EMA data is imported to Artisynth to run the simulation of the inverse tongue model. Initial attempts revealed that the span of the original tongue data was too large to be handled by the model, causing the simulations to become unstable and stop. Investigations revealed that this was caused by one muscle being saturated for a long time when the model tried to reach the target position. To avoid the instability in simulations, the span of the tongue movement was decreased to 70% of the original data. This scaling was determined experimentally based on muscle activation levels, and a future, more elaborate study should instead set the scaling based on differences in articulatory space for the subject measured with EMA and that used for the tongue model.

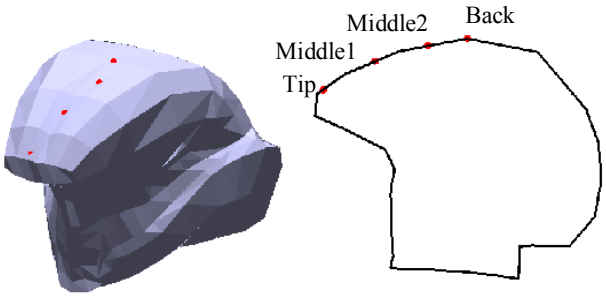

Figure 3. Virtual sensor positions on Artisynth tongue model

5. Results

We exemplify the results obtained with the /ak:a/ utterance. First, we consider the target and model positions to determine how well the Artisynth inverse tongue model follows the target trajectory. We then consider the muscle activation used to create the trajectories. In the following discussion and figures, X, Y, and Z directions correspond to the coronal, sagittal and transverse planes, respectively.

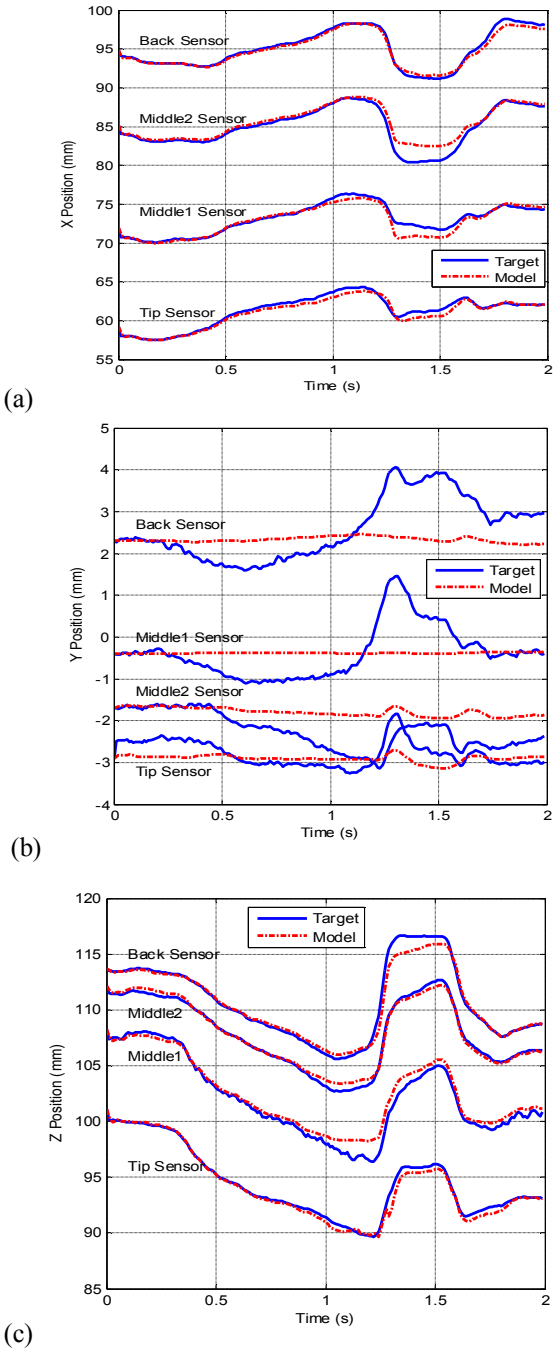

Figure 4. Position of target and model for the four sensors in the (a) X-, (b) Y- and (c) Z-directions

5.1 Target and Model position

Figure 4 compares the four target and corresponding model positions in three different directions. It is apparent from Figure 4 (a) and (c) that the model is able to follow the target points with good accuracy in the midsagittal plane (X- and Z-directions) during the initial and final /a/ (up to 1.2 s and after 1.6 s). During /k:/, at time 1.2-1.6 s, the model recreates the same midsagittal movement, but the error increases. This could be caused by two different factors. Firstly, as the tongue moves to articulate the occlusion in /k:/, the velocity increases and the model may have more difficulties following the targets, even if the velocity is not substantially larger than during the /a/. Secondly, system latency may introduce a delay in decreasing the prediction error for larger position changes: in general, the model trajectory is slightly smoother than the input. The performance for more varied phonetic input and the cause of the increase in prediction error will be investigated further in future studies.

Figure 4 (c) shows that the model completely fails to follow the target positions in the lateral (Y-direction). The span of target position in the Y-direction is however less than 2 mm, and investigations of the four target positions indicate that the variations are often in opposite directions: when one target point moves to the right, another moves to the left. It is apparent that the tongue model structure cannot account for this kind of skew movements in the Y-direction.

Since lateral tongue movements have little impact on the acoustic output, it may indeed be reasonable to discard these small variations in the Y-direction and perform the inversion solely in the midsagittal plane, using X- and Z-data.

A video of resynthesized tongue movements is available in EUNISON YouTube channel [11].

5.2 Muscle activation

Figure 5 depicts the audio waveform, sensors position error and muscle activation for utterance /ak:a/, where the sensors position error is the Euclidean distance between the target and model in 3D coordinates. According to figure 5, none of the muscles is active during the initial silent period and the tongue is in rest position. The muscle activations then start slightly before the acoustic output. The most active muscles during the production of the vowel /a/ are GGA and STY. When the tongue is moved to articulate /k:/ at 1.2 s, these muscles relax, and GGP becomes the most active, briefly seconded by TRANS. During the final /a/ the GGP activity goes down, and GGA becomes active again, but with weaker activation. It can be noted that slight muscle activation persists after the production of the final /a/, but this is natural, since the utterance was extracted from a longer sequence of vowel-consonant-vowel combinations.

We have compared muscle activation for the phoneme sequences /ɑ:i:/, /ɑ:u:/, /i:u:/ and /ak:a/, and even if it is not possible to claim that the inversion procedure has determined *the actual* muscle activation used by the subject, the results nevertheless appear plausible. In a pre-study [12] to this paper, the second author compared the dominating muscles in the Artisynth model for the cardinal vowels with previous studies, and found a general good correspondence, as summarized in Table 1.

Table 1. Muscle activation, in order of importance, found by the Artisynth model in a pre-study [11], compared to literature references. Matches indicated in bold.

	Artisynth inversion	**Literature sources**
/*a:*/	**GGa**, GGm, SL, **SG**, **HG**	**HG**, **GGa [13] or SG [7]**
/i:/	**GGp**, IL, **MH**	**GGp**, **MH** and GGa [13]
/u:/	**GGp**, **SG**, **GGa**, **MH**	**SG**, **GGp**, **GGa**, HG, **MH [13]**

In general most of the muscles, which could be expected to be active, based on earlier studies [7][13], were also activated in the inversion, even if the order of importance differed to some extent, and some mismatches were found. This could be due to differences between speakers and corpora, and not necessarily to unexpected behaviour of the inversion model.

6. Conclusions and Discussion

In this preliminary work, we present our simulations and observations of studying muscle activation using the biomechanical tongue model in Artisynth in combination with EMA data. This study indicates that this combination could be used as a powerful tool for phonetic research, since the tongue movements observed in the EMA measurements can, to a large extent, be replicated with the model, by activating muscles in a plausible manner. The study has further identified several issues for future studies, such as increased prediction error for more rapid tongue movements and the effect of speaker differences in articulatory space. It should further be noted that the present study focuses solely on tongue movements and disregards movements of the jaw. Since it is possible to perfectly produce the studied sequences with a fixed jaw (c.f. bite-block experiments, e.g. [14]), the current results do have bearing on natural speech production. Future studies should nevertheless include the jaw and instead consider the tongue movements that are independent of the jaw.

7. Acknowledgments

This research has been supported by EU-FET grant EUNISON 308874. The EMA data used in this paper was collected in collaboration with Slim Ouni at Loria (Laboratoire Lorrain de Recherche en Informatique et ses Application), Nancy, as part of the ASPI project (aspi.loria.fr)

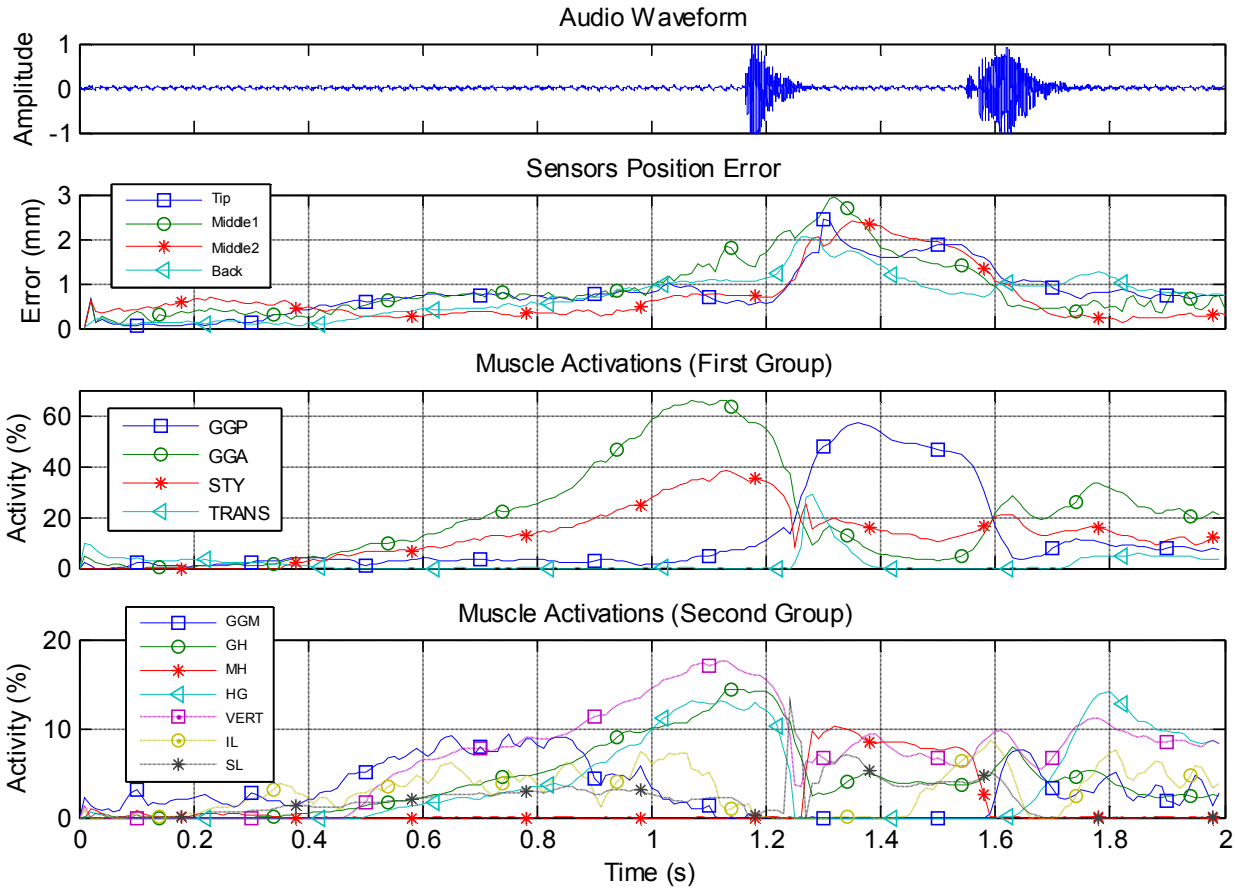

Figure 5. Audio waveform, sensors position errors and muscle activations for utterance /aka/

References

[1] J. S. Perkell, M. H. Cohen, M. a Svirsky, M. L. Matthies, I. Garabieta, and M. T. Jackson, "Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements.," J. Acoust. Soc. Am., vol. 92, no. 6, pp. 3078–96, Dec. 1992.

[2] A. Zierdt and P. Hoole, "Extracting tongues from moving heads," Proc. 5th Speech Production Seminar, 2000.

[3] I. Steiner and S. Ouni, "Towards an articulatory tongue model using 3D EMA," 9th Int. Seminar on Speech Production, vol. 11, pp. 147–154, 2011..

[4] Badin, P., Baricchi, E. & Vilain, A. (1997). "Determining tongue articulation: from discrete fleshpoints to continuous shadow". In 5th EuroSpeech Conference, vol. 1, pp. 47-50. Rhodos, Greece, September 1997.

[5] A. Toutios and S. Maeda, "Articulatory VCV Synthesis from EMA Data.," *INTERSPEECH*, 2012.

[6] J.M. Gerard, R. Wilhelms-Tricarico, Y. Payan and P. Perrier, "A 3D biomechanical tongue model to simulate speech movements.," Archives of Physiology and Biochemistry vol. 111, pp. 7, 2003.

[7] Q. Fang, S. Fujita, X. Lu, and J. Dang, "A model-based investigation of activations of the tongue muscles in vowel production," Acoust. Sci. Technol., vol. 30, no. 4, pp. 277–287, 2009.

[8] I. Stavness, J. E. Lloyd, and S. Fels, "Automatic prediction of tongue muscle activations using a finite element model.," *J. Biomech.*, vol. 45, no. 16, pp. 2841–8, Nov. 2012.

[9] Carstens AG500 articulograph, http://www.articulograph.de/

[10] J. E. Lloyd, I. Stavness, "Artisynth Description and design overview," 2011. www.magic.ubc.ca/artisynth/artisynth/doc/artisynth.pdf

[11] EUNISON YouTube channel, https://www.youtube.com/user/eunisonFet

[12] Nilsson, I. "Inverse Modelling of Biomechanical Tongue Control. From Articulation Data to Muscle Activation". Bachelor thesis, KTH Royal Institute of Technology, Stockholm, Sweden, 2014.

[13] Baer, T., Alfonso, P.J. & Honda, K. (1988). Electromyography of the tongue muscles during vowels in /əpvp/ environment. *Annual Bulletin, Research Institute Of Logopedics And Phoniatrics*, pp. 7-19.

[14] Lindblom, B., Lubker, J. & McAllister, R. (1977): "Compensatory articulation and the modeling of normal speech production behavior". In Proceedings from Symposium on Articulatory modeling and phonetics. Grenoble, 1977.