MULTIMODAL SPEECH ANIMATION FROM ELECTROMAGNETIC ARTICULOGRAPHY DATA

Guillaume Gibert[1,2,3,4], Virginie Attina[4], Mark Tiede[5,6], Rikke Bundgaard-Nielsen[4], Christian Kroos[4], Benjawan Kasisopa[4], Eric Vatikiotis-Bateson[7], Catherine T. Best[4,5]

[1]Inserm, U846, 18 avenue Doyen Lépine, 69500 Bron, France
[2]Stem Cell and Brain Research Institute, 69500 Bron, France
[3]Université de Lyon, Université Lyon 1, 69003, Lyon, France
[4]Marcs Institute, University of Western Sydney, Locked Bag 1797, Penrith NSW 2751, Australia
[5]Haskins Laboratories, New Haven CT, U.S.A. 06510
[6]Speech Comm. Group, R.L.E. MIT
[7]Cognitive Systems & Linguistics, U. British Columbia

ABSTRACT

Virtual humans have become part of our everyday life (movies, internet, and computer games). Even though they are more and more realistic, their speech capabilities are, most of the time, limited and not coherent and/or not synchronous with the corresponding acoustic signal. We describe a method to convert a virtual human avatar (animated through key frames and interpolation) into a more naturalistic talking head. Speech-capabilities were added to the avatar using real speech production data. Electromagnetic articulography (EMA) data provided lip, jaw and tongue trajectories of a speaker involved in face to face communication. An articulatory model driving jaw, lip and tongue movements was built. Constraining the key frame values, a corresponding high definition tongue articulatory model was developed. The resulting avatar was able to produce visible and partly occluded facial speech movements coherent and synchronous with the acoustic signal.

Index Terms— Speech production, Talking head, ElectroMagnetic Articulography (EMA), Augmented speech

1. INTRODUCTION

Humans employ speech reading commonly in adverse listening conditions and in general to facilitate speech perception [1]. The ability to visually obtain phonetic information depends on seeing facial movements that are produced by the speech articulators: mainly by the lips and the jaw and to some extent by the larynx and the tongue. These movements have been shown to be highly correlated with speech acoustics [2].

Although the tongue is not visible most of the time, its movements provide useful information for visual speech perception as shown in [3] where perceivers performed better with point-light displays including additional dots on the tongue and the teeth than with displays with 'lips only' dots.

Virtual humans are part of our everyday life. They can be found in 3D movies, on the internet as helping agents and in computer games. Speech capabilities of these avatars are in general very poor. Lip, jaw, and tongue movements are often not coherent and/or not synchronous with the corresponding acoustic signal. These flaws may lead to misperceptions such as those encountered during the perception of incongruent auditory-visual speech stimuli [4]. Because of theses issues, hearing impaired people and second language learners cannot make full use of these technologies. This remains the case, even though talking heads have been specifically developed to help hearing impaired people [5] and to teach second language learners [6]. However, accurate 3D tongue models have been included in talking heads. These models were obtained by Magnetic Resonance Imaging (MRI) [7, 8] and animated by EMA data [9]. Therefore, it is now possible to incorporate natural, realistic articulation patterns into avatars.

We propose a method to transform an existing avatar animated by linear interpolation between key frames into a talking head. An electromagnetic articulography (EMA) system was used to record a person during face to face communication. Lip, jaw and tongue trajectories were recorded with this system. We built a low definition (LD) articulatory model by decomposing each speech articulator's movements separately using guided PCA. This LD model was then used to build a high definition (HD) tongue model from the available set of key frames. EMA sensor positions were inversed and articulatory parameters were computed. Applying the model enabled the avatar to produce realistic speech movements driven by articulatory parameters derived from EMA data.

2. METHODS

2.1. Setup

An EMA system (WAVE, Northern Digital Inc.) was used to record the position and orientation of sensors attached to the gum, lips and tongue during a speech experiment at the Marcs Institute Speech Production Lab (MISPL), University of Western Sydney. Its field transmitter emits an electromagnetic field and signals transduced in small sensors (3 mm²) within the field are resolved into spatial positions. The optimal measurements were within a 30 cm virtual cube oriented to the transmitter unit. The system delivered three spatial (x,y,z) and two angular (azimuth, elevation) measurements per sample and per sensor at 100 Hz. The accuracy of the tracking system has been previously assessed and validated for speech research [10].

In the experiment, the position and orientation of nine sensors were recorded at 100 Hz. Two sensors were attached to the lips: on the upper lip (UL) and on the lower lip (LL). Using dental glue, two sensors were glued on the incisors: UI on the gumline of the upper incisor and Jaw on the gumline of the lower incisor. Three sensors were glued on the tongue: TT on the tongue tip, TB on the tongue body and TD on the tongue dorsum. Two sensors (LM and LR) were also attached to the left and right mastoid processes to correct for head movement. The positions of the EMA sensors are shown in Figure 1. The audio signal (mono, 22.05 kHz, 16 bits) was recorded synchronously by the EMA system.

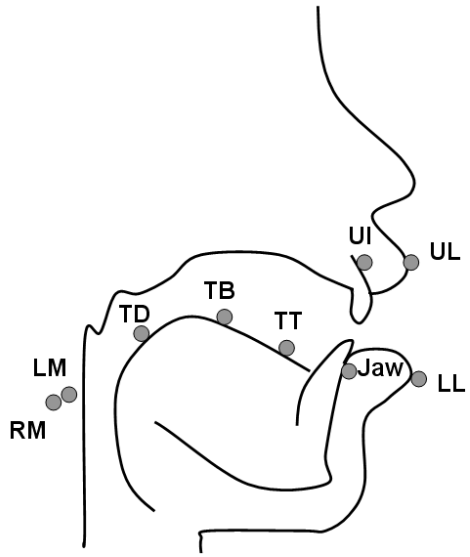


Figure 1: Schematic midsagittal view of the speaker showing the position of the EMA sensors. Two sensors were attached on the lips (UL, LL), two sensors were positioned on the incisors (Jaw, UI), three sensors were glued on the tongue (TT, TB, TD) and two sensors were attached to the left and right mastoids (LM, RM, respectively).

2.2. Participant

Two native speakers of American English participated in this recording. We used the male speaker's recordings here.

2.3. Design and procedure

The data were collected in a joint experiment reported in [11]. The data presented in this paper represent a subset of the whole recording.

One speaker was seated close to the WAVE transmitter and facing another American English female speaker 2 meters away. Speakers were instructed to have a conversation on any topic of their choice, after which they were asked to produce tongue-twister sequences simultaneously (speech competition) under different speech rates [11]. Five dyadic interactions were recorded, each one having duration of 120s. Therefore, the total number of frames was 60000.

3. EMA DATA MODELLING

Head movement (translations and rotations) was estimated and corrected using the landmarks positioned on the upper incisor (*UI*) and on the mastoid processes (*LM, RM*). These sensors were not affected by speech articulator movements. An articulatory model was built using the method proposed by [5, 12]. A pruning step (simple vector quantization) was applied to remove the frames in which sensor positions were similar (Euclidian distance < 0.9 mm). It conditioned the data before building the statistical models. Then, the contribution of the different speech articulators (jaw, lips and tongue) was iteratively subtracted. This subtraction consisted of an iterative application of Principal Component Analysis (PCA) on subsets of landmarks. The procedure extracted nine articulatory parameters:

- Jaw opening (**jaw1**) using PCA on the *Jaw* position sensor values (39.4% of the global variance);
- Jaw rotation (**jaw2**) using PCA on the residual *Jaw* position sensor values (1% of the global variance);
- Lip rounding (**lips1**) using PCA on the residual lip (*UL, LL*) position values (2.7% of the global variance);
- Lip closing (**lips2**) using PCA on the residual lower lip (*LL*) position values (1.5% of the global variance);
- Lip raising (**lips3**) using PCA on the residual upper lip (*UL*) position values (0.7% of the global variance);
- Tongue front-back movement (**tongue1**) using PCA on the residual tongue (*TB, TD*) position values (16.8% of the global variance);

- Tongue flattening-bunching movement (**tongue2**) using PCA on the residual tongue (*TB, TD*) position values (4.7% of the global variance);
- Tongue tip vertical movement (**tongue3**) using PCA on the residual tongue (*TT*) position values (20.1% of the global variance);
- Tongue tip horizontal movement (**tongue4**) using PCA on the residual tongue (*TT*) position values (7.1% of the global variance);

More than 94% of the global variance was explained by these 9 articulatory parameters. This modeling procedure had two aims: first, to extract meaningful parameters controlling an elementary articulator, and second, to remove artifacts and measuring noise. This articulatory model will be referred to hereafter as the Low Dimension (LD) model.

4. ANIMATION

4.1. Avatar

The avatar used in this study was a representation of the Australian performance artist Stelarc. This 3D model was originally driven by a set of key frames controlling the visible and partially occluded speech facial articulators such as lips, jaw, and tongue. The full animation was originally created by linear interpolations between those key frames. Unfortunately, linear interpolations do not accurately replicate speech articulator movements. This is one of the reasons why we developed a new animation method.

4.2. Face parameters creation

Selected key frames (from the original model) were used to create articulatory parameters for driving the avatar. The vertex coordinates of the neutral pose were subtracted from the vertex coordinates of each key frame. The resulting variation between these positions was then variance-normalized and set to vary between 0 and +3. Synthetic articulatory parameters controlling the jaw (and the mandible) (**jaw1**) and the lips (**lips1**, **lips2**, **and lips3**) were created. These parameters corresponded to the facial articulatory parameters derived from EMA data. Note that no parameter corresponding to **jaw2** was found in the available key frames. Since this parameter recovered only 1% of the global variance in the EMA data, it was not included in the final set of synthetic articulatory parameters for the animation.

4.3. Tongue parameters creation

Because tongue key frames were not related to any speech articulation in the original animation, but only to meaningless geometric variation, an alternative method was designed. Each tongue sensor from EMA data was associated with a specific vertex of the 3D tongue mesh of the original face model. For each sample of the quantized EMA database, tongue postures were determined by estimating the best linear mixture of weighted key frames that minimized the distance between the EMA tongue sensor positions and the corresponding tongue mesh vertices. The least square estimation of the vector of weights α was simply performed by:

$$\hat{\alpha} = \underset{\alpha \in [-10;10]^N}{\operatorname{argmin}} \left\| \sum_{i}^{N} \alpha_i P3D_{K_i} - P3D_{EMA} \right\|_2$$

where $P3D_{K_i}$ corresponded to the position of the three selected vertices of the 3D tongue mesh for the key frame K_i, α_i corresponded to the weights applied to the key frame K_i, and $P3D_{EMA}$ corresponded to the position of the three EMA sensors TD, TB and TT. The number of key frames available in the original model was N = 9. The values of each weight α_i were limited to [-10; 10]. Examples of configurations found in the EMA database and the corresponding constrained 3D tongue mesh can be visualized in Figure 2.

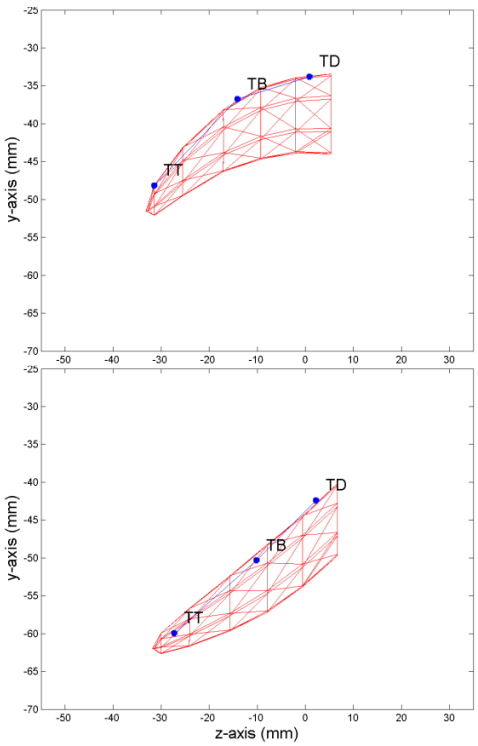
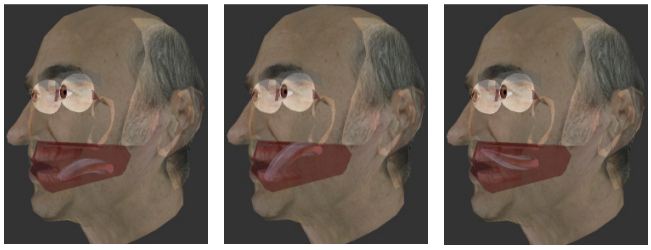


Figure 2: Two tongue configurations (midsagittal view) from the quantized EMA database (TD, TB and TT sensor positions in blue) and the corresponding constrained 3D tongue mesh (red mesh).

After this step, a quantized database of HD tongue postures was created. For all the configurations of the EMA database, corresponding constrained 3D tongue mesh was available. The reconstruction error computed as the Euclidian distance between the EMA sensor (TD, TB and TT) positions and the specific vertices of the 3D tongue mesh was M = 7.07 mm and SD = 6.94 mm.

The same procedure as described in section 3 was used to build a high dimension tongue articulatory model using the database of 3D tongue postures in addition to the EMA database. Finally, the HD tongue model was controlled by 5 articulatory parameters (as described in [13]): jaw height/opening (**jaw1**), tongue front-back (**tongue1**), tongue flattening-bunching (**tongue2**), tongue tip vertical (**tongue3**) and tongue tip horizontal (**tongue4**). Examples of the maximum variation of key articulatory parameters are shown in Figure 3.



(a) *Tongue1* (b) *Tongue2* (c) *Tongue3*

Figure 3: Examples of the maximum variation (one direction) of some articulatory parameters driving the tongue: (a) *Tongue1* displays the posterior extent of a tongue front-back movement; (b) *Tongue2* displays the peak of a flattening-bunching movement; (c) *Tongue3* displays the peak of a tongue tip vertical movement.

4.4. Articulatory parameters inversion

For each frame of the recorded sequences, articulatory parameters were estimated in order to minimize the distance between the actual sensor positions and the reconstructed ones. The least square estimation of the vector of articulatory parameters β was performed by:

$$\hat{\beta} = \underset{\beta \in [-3;3]^M}{\operatorname{argmin}} \| m_{Face} + \beta eigv_{Face} - P3D_{EMA} \|_2$$

where $P3D_{EMA}$ corresponded to the position of the EMA sensors after subtraction of the rigid head motion, m_{Face} corresponded to the mean face configuration of the articulatory model, $eigv_{Face}$ corresponded to the matrix of eigenvectors of the articulatory model, β corresponded to the articulatory parameter values. The number M of articulatory parameters driving the tongue was 5 (**jaw1**, **tongue1**, **tongue2**, **tongue3**, **tongue4**). The values of each articulatory parameter were limited to [-3; 3].

4.5. Animation

Because there was a complete correspondence between the LD (from EMA data) and the HD (from the avatar) articulatory models (except for jaw2), the articulatory parameter values derived from EMA data could be used directly to animate the avatar. Therefore, the avatar can repeat what the speaker said using only the EMA data to drive his jaw, lip and tongue movements. Examples of animation can be viewed at:
http://swooz.free.fr/download/gibert_etal_eusipco2012.wmv
An example of variation of the tongue articulatory parameters over time for a sentence pronounced by the speaker is presented in Figure 4. The articulatory parameters did not vary linearly between articulatory targets. The animation module of the original avatar could not generate these nonlinear trajectories. On the contrary, the new animation technique can generate series of 3D postures corresponding to these nonlinear trajectories.

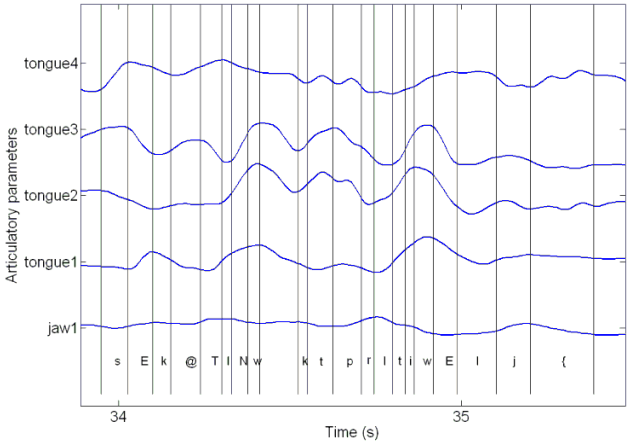


Figure 4: Variation of the tongue articulatory parameters for the sentence "second thing worked pretty well yeah".

5. CONCLUSIONS & PERSPECTIVES

A method to transform an avatar with generic tongue model and animation by key frames into a talking head that displays naturalistic tongue, jaw and lip motions was described. First, real speech articulation data were used to build an LD articulatory model. It consisted of nine articulatory parameters driving speech articulators: jaw, lips and tongue. An HD articulatory model was then created by transforming selected key frames into articulatory parameters for jaw and lips and by constraining the generic tongue model to the LD tongue model. Real articulatory data together with the acoustic signal were used to steer the talking head.

A limitation of the proposed method came from the small numbers of sensors used. It was not possible to

process the face and tongue models in the same way. Thus, a promising next step would be to use synchronous recordings from WAVE and Optotrak Certus (Northern Digital Inc.) systems. With this setup, a large number of sensors could be placed on the speaker's face and tongue. Consequently, the method to derive the tongue model presented in this paper could be used to create a more accurate HD facial articulatory model. The parameter **jaw2** may be easily created with this procedure.

The proposed method could be applied to modify existing avatars that are not able to produce correct speech movements. This would allow hearing impaired people and second language learners to effectively utilize a larger number of virtual agent applications. An evaluation of the method will be performed to assess the gain in intelligibility, for instance, through a speech in noise perception experiment.

6. ACKNOWLEDGEMENTS

We thank Steve Fazio for his technical support during the recording and Kirk Olsen for manually segmenting the audio files. This work was supported by ARC Human Communication Science Network (RN0460284), by the Thinking Head project [14], a special initiative scheme of the Australian Research Council and the National Health and Medical Research Council (TS0669874), by Marcs Institute, University of Western Sydney and by the SWoOZ project (ANR 11 PDOC 019 01).

7. REFERENCES

[1] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *Journal of the Acoustical Society of America,* vol. 26, pp. 212-215, 1954.

[2] H. C. Yehia, T. Kuratate, and E. Vatikiotis-Bateson, "Linking facial animation, head motion and speech acoustics," *Journal of Phonetics,* vol. 30, pp. 555-568, 2002.

[3] L. D. Rosenblum, J. A. Johnson, and H. M. Saldana, "Point-light facial displays enhance comprehension of speech in noise," *Journal of Speech and Hearing Research,* vol. 39, pp. 1159-1170, 1996.

[4] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature,* vol. 264, pp. 746-8, Dec 23-30 1976.

[5] G. Gibert, G. Bailly, D. Beautemps, F. Elisei, and R. Brun, "Analysis and synthesis of the three-dimensional movements of the head, face, and hand of a speaker using cued speech," *Journal of the Acoustical Society of America,* vol. 118, pp. 1144-1153, Aug 2005.

[6] O. Engwall, "Can audio-visual instructions help learners improve their articulation? An ultrasound study of short term changes," in *Interspeech 2008*, Brisbane, Australia, 2008, pp. 2631-2634.

[7] P. Badin, G. Bailly, L. Reveret, 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, pp. 533-553, 2002.

[8] O. Engwall, "A 3D tongue model based on MRI data," in *International Conference on Spoken Language Processing*, Beijing, China, 2000, pp. 901-904.

[9] O. Engwall, "Combining MRI, EMA and EPG measurements in a three-dimensional tongue model," *Speech Communication,* vol. 41, pp. 303-329, Oct 2003.

[10] J. J. Berry, "Accuracy of the NDI Wave Speech Research System," *J Speech Lang Hear Res,* vol. 54, pp. 1295-1301, October 1, 2011.

[11] M. Tiede, R. Bundgaard-Nielsen, C. Kroos, G. Gibert, V. Attina, B. Kasisopa, E. Vatikiotis-Bateson, and C. Best, "Speech articulator movements recorded from facing talkers using two electromagnetic articulometer systems simultaneously," *The Journal of the Acoustical Society of America,* vol. 128, pp. 2459-2459.

[12] L. Reveret, G. Bailly, and P. Badin, "MOTHER: A new generation of talking heads providing a flexible articulatory control for video-realistic speech animation," in *6th Int. Conference of Spoken Language Processing, ICSLP'2000*, Beijing, China, 2000.

[13] P. Badin and A. Serrurier, "Three-dimensional linear modeling of tongue: Articulatory data and models," in *7th International Seminar on Speech Production*, D. D. R. L. H.C. Yehia, Ed. Belo Horizonte, Brazil, 2006, pp. 395-402.

[14] D. Burnham, R. Dale, K. Stevens, D. Powers, C. Davis, J. Buchholz, K. Kuratate, J. Kim, G. Paine, C. Kitamura, M. Wagner, S. Möller, A. Black, T. Schultz, and H. Bothe, "From Talking Heads to Thinking Heads: A Research Platform for Human Communication Science," ARC/NH&MRC Special Initiatives, TS0669874, 2006-2011.