An Audiovisual Talking Head for Augmented Speech Generation: Models and Animations Based on a Real Speaker's Articulatory Data

Pierre Badin, Frédéric Elisei, Gérard Bailly, and Yuliya Tarabalka

GIPSA-lab / DPC, UMR 5216 CNRS – INPG – UJF – Université Stendhal, Grenoble, 961 rue de la Houille Blanche, BP 46, F-38402 Saint Martin d'Hères Cedex, France
{Pierre.Badin,Frederic.Elisei,Gerard.Bailly,Yuliya Tarabalka}
@gipsa-lab.inpg.fr

Abstract. We present a methodology developed to derive three-dimensional models of speech articulators from volume MRI and multiple view video images acquired on one speaker. Linear component analysis is used to model these highly deformable articulators as the weighted sum of a small number of basic shapes corresponding to the articulators' degrees of freedom for speech. These models are assembled into an audiovisual talking head that can produce augmented audiovisual speech, i.e. can display usually non visible articulators such as tongue or velum. The talking head is then animated by recovering its control parameters by inversion from the coordinates of a small number of points of the articulators of the same speaker tracked by Electro-Magnetic Articulography. The augmented speech produced points the way to promising applications in the domain of speech therapy for speech retarded children, perception and production rehabilitation of hearing impaired children, and pronunciation training for second language learners.

1 Introduction

The importance of the visual speech input in language acquisition has been well documented by [1] in a review on language acquisition in blind children, with data showing that blind children have difficulty in learning an easy-to-see and hard-to-hear contrast such as [m] vs. [n]. This importance is also demonstrated by data showing the predominance of bilabials at the first stage of language acquisition ([2]), predominance which is reinforced in hearing impaired children ([3]), but less clear in blind ones ([4]). More generally, the importance of vision in speech perception has been demonstrated in a large number of studies: [5], and more recently [6], among others, have quantified the gain in speech intelligibility provided by the vision of lips and face in comparison with the sole acoustic signal (e.g. up to 35% for a signal to noise ratio of –9 dB). In addition, it is known that human beings possess, up to a certain level, a general *articulatory awareness* skill, i.e. the ability to know the shape and position of one's own articulators (as measured e.g. by [7] for internal speech articulators such as tongue).

F.J. Perales and R.B. Fisher (Eds.): AMDO 2008, LNCS 5098, pp. 132–143, 2008.

Virtual audiovisual talking heads can produce *augmented* audiovisual speech, i.e. can display not only usually visible articulators such as lips but also usually *non* visible articulators such as tongue or velum. These capabilities of augmented display present thus a potentially high interest for applications such as therapy for speech retarded children, rehabilitation of perception and production for hearing impaired children ([8]), or pronunciation training for second language learners ([9]).

In the framework of speech production studies, we develop models of speech articulators ([10], [11], [12], [13]). The present article describes the methods used to build these models from data acquired from a real speaker, how these models are integrated in a virtual audiovisual talking head, and how this talking head can be animated from recordings of the speaker by motion capture.

2 Articulatory Data and Models for Speech Articulators

The speech apparatus is made of a number of articulators, some of them rigid, such as the jaw or the hyoid bone, some of them highly deformable, such as the tongue or the lips. These articulators, in conjunction with other less deformable structures such as the pharyngeal wall or the cheek walls, shape a highly deformable tube – the vocal tract – that extends from the vocal folds to the lips. As mentioned by [14], each articulator may be made of large number of neuromuscular components which offer a potentially huge dimensionality and which must be functionally coupled in order to produce the relatively simple gestures associated with speech production tasks. Modelling these articulators consists thus in extracting their necessarily small number of *degrees of freedom* (henceforth DoF) – or *components* – from appropriate articulatory data acquired from a human subject.

2.1 Modelling Principles

We have developed the articulators' models according to a *speaker-oriented data-driven linear* modelling approach that we summarise in the present section (see details in [10]). Each component is specified by the limited set of movements that it can execute independently of the other components. Rather than defining these components for each articulator *a priori*, and attempting to fit them to real speakers' movements *a posteriori*, we extract them from a corpus of articulations (corresponding to phonemes) representative of the speech production capabilities of a given speaker. This *speaker-oriented* approach allows a rigorous evaluation of the models, by direct comparison with ground-truth data. Besides, using a single speaker avoids merging the anatomical characteristics and the control strategies that can vary fairly much among speakers.

In the framework of our *linear* modelling approach, the geometry of the various non-rigid articulators is modelled as the weighted sum of a small number of basic shapes, associated with the DoFs, which constitutes a minimal basis for the space of articulations. Whereas data-driven models are classically built using *Principal Component Analysis (PCA)*, our models are built using a so-called *guided PCA* where *a priori* knowledge is introduced during the linear decomposition. The weights constitute the *articulatory control parameters* associated with the components: a given set of values of these parameters produces a given shape of the articulators. We ensure that each component is *linearly* uncorrelated with the other components over the set of tasks considered. Our approach aims at finding a compromise between two possibly conflicting criteria: (1) reducing the number of DoFs of the articulators as much as possible by exploiting correlations in the articulatory data; (2) maintaining *biomechanical likelihood*, i.e. making sure that the DoFs are not related to the control strategies actually used by the speaker during the task but are really associated with movements that are plausible from the viewpoint of biomechanics.

Each linear DoF is *iteratively* determined by means of a mixture of PCA applied to carefully chosen data subsets of the articulators' shapes and of multiple regression of the data against control parameters either arbitrarily imposed such as jaw height or determined by the preceding PCA. Note that the solution of this type of linear decomposition is not unique in general: while PCA delivers optimal factors explaining the maximum of data variance with a minimum number of components, our approach allows some freedom to decide the nature and distribution of the variance explained by the components (for instance to make them more interpretable in terms of control), at the cost of a sub-optimal variance explanation and of weak correlation between components.

2.2 Determination of the Articulators' Shapes from Various Type of Data

The complex geometry of the various speech articulators / cavities, e.g. tongue tip, velopharyngeal port, is determined as three-dimensional surfaces from data obtained using different setups. Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) are used for internal articulators, while multiple view video recordings are used for the lips and face. A corpus has been designed to cover the maximal range of French articulations that the speaker can utter: it consists of the oral and nasal vowels [a ɛ e i y u o ø ɔ œ ɑ̃ ɛ̃ œ̃ ɔ̃], of the consonants [p t k f s ʃ m n ʁ l] in three symmetrical contexts [a i u], and of a rest and a pre-phonatory articulations.

Internal articulators. The complete three-dimensional surface representations of the internal articulators that are needed to build the three-dimensional articulators models can only be obtained from images provided by medical imaging systems such as CT and MRI systems. A stack of axial images of the head of the speaker at rest was made by CT, to serve as a reference. These CT images, that provide a good contrast between bones, soft tissues and air, are used to locate bony structures and to determine accurately their shapes. Stacks of sagittal MR images were recorded for the speaker sustaining artificially each of the articulations of the corpus during about 35 sec. A set of 25 sagittal images was obtained for each articulation. These images provide a good contrast between soft tissues and air, and also within soft tissues, but do not image clearly the bones. Due to the complexity of the contours of the various articulators, to the relatively low resolution of the images (about 1 pixel / mm), and to the need of an accurate reconstruction of the articulators, the extraction of contours has been performed manually, plane by plane (see e.g. Fig. 1). In order to improve the determination of the articulator outline in some regions, transverse images were created by re-slicing the initial stack of images in appropriate planes (see e.g. Fig. 2).

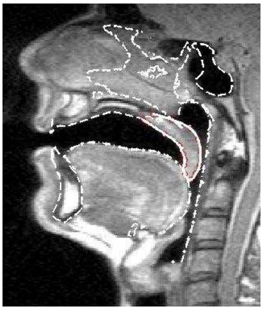


Fig. 1. Example of edited contours in the midsagittal plane on top of an MRI image

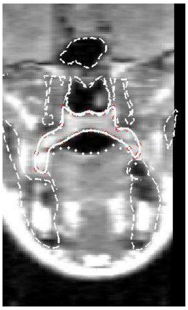


Fig. 2. Idem in a transverse plane

The contours of the rigid bony structures involved in the vocal tract (jaw, hard palate, nasal passages, nostrils and various paranasal sinuses) have been manually registered as planar B-spline curves controlled by a limited number of points from CT images in planes with appropriate orientations. The set of all points forming these 2D contours was then expanded into a common reference 3D coordinate system attached to the skull of the speaker. These 3D points were finally processed through a 3D meshing software (Geometrica Research Group at INRIA, http://cgal.inria.fr/Reconstruction) to form a 3D surface meshing based on triangles (see e.g. Fig. 4 for an illustration of the jaw surface).

In order to compensate for possible slight position changes of the speaker's head between the recordings of the various articulations, each stack is aligned with a common reference 3D coordinate system using an appropriate 3D transformation. This transformation, which corresponds to the six degrees of freedom of a solid object, henceforth *3D rototranslation*, is obtained by aligning the rigid structures extracted from CT images (hard palate, nasal passages, paranasal sinuses) with the corresponding ones in the MR images stack. The same procedure is also applied to the jaw for each articulation, to determine its relative position in relation to the fixed rigid structures.

The determination of the deformable structures (tongue, velum, nasopharyngeal wall) is achieved in much the same way as for the rigid structures, but from the MR images of each articulation. The sets of all 2D planar contours edited from the MRI images are expanded into the 3D coordinate system to form the primary 3D description of the deformable articulators' shape. As linear analysis methods such as PCA require each observation to bear on the same number of variables, it is needed to represent the shape of each articulator with the same number of points for all the articulations of the corpus. Such a suitable geometric representation was obtained in the following way each deformable articulator: a unique generic 3D surface mesh, made of triangles, was defined for the articulator, and was fitted by elastic deformation to each of the articulator's 3D primary shapes for all articulations of the corpus.

We finally obtained, the shapes of each articulator as 3D surfaces defined in terms of triangular meshes having the same number of vertices for each of the 46 articulations of the corpus: the tongue is made of 1640 vertices (the RMS reconstruction error over the corpus is 0.06 cm); the velum has 5239 vertices (RMS = 0.06 cm). This forms the basis for the articulatory modelling, as illustrated further.

Visible articulators. Measurements of visible articulators, essentially face and lips surfaces, can be obtained from video recordings. About 250 coloured markers are attached to the speaker's face (cf. [10] or [12]). The use of multiple synchronous and calibrated cameras allows determining the 3D coordinates of these markers with accuracy better than one millimetre. The surface of the face is represented by a mesh of about 450 triangles of which vertices are hooked to these points, as illustrated in Fig. 3. Besides, the lips are described by means of a generic mesh. This mesh, which is speaker-independent, is deformed and fitted to the images by an expert, using the multiple views to capture the lips external contours and to model their visible surface.

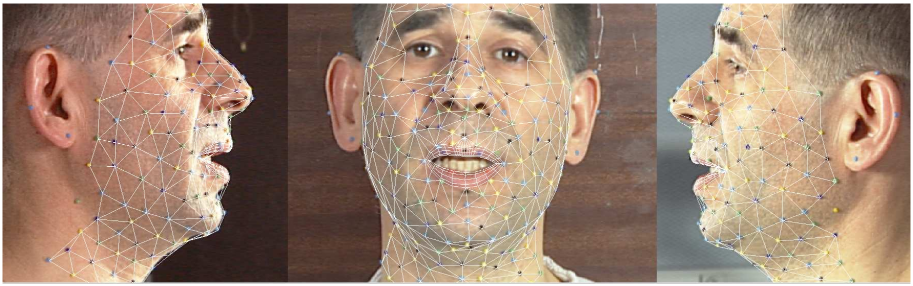


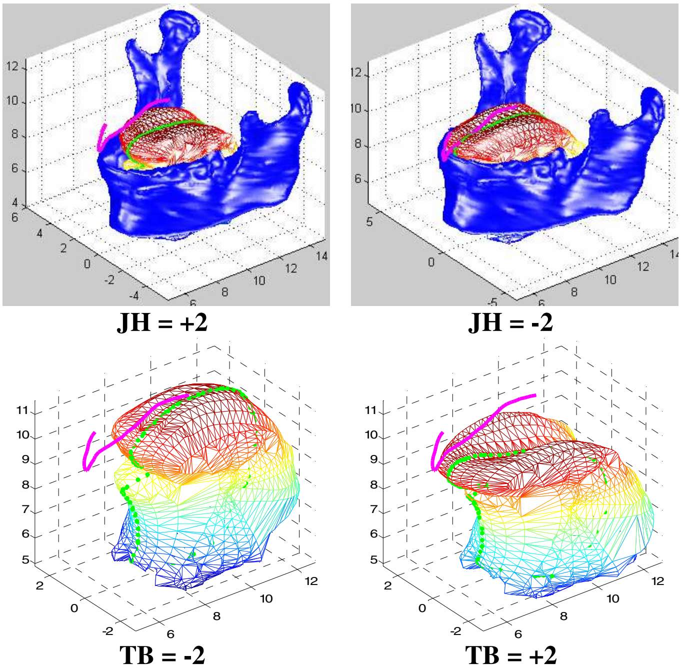
Fig. 3. Example of view of a subject's head produced by three synchronous and calibrated cameras, with superimposed lips and face meshes

2.3 Articulatory Models of Jaw, Tongue, Velum, Lips, and Face

The modelling principles described in section 2.1 have been applied to the 3D surface meshes of the various speech articulators. Details can be found in [11] for the tongue, in [13] for the velum, and in [12] for the lips and face.

Jaw displacement model. The six parameters of the 3D rototranslation of the jaw rigid body are very much correlated with the vertical and horizontal coordinates of the upper edge of the lower incisor, called *jaw height*, and *jaw advance*. The centred and normalised versions of these two parameters, *JH* and *JA*, are therefore used as control parameters of the jaw 3D rototranslation (see Fig. 4 for an illustration of the movements of the jaw associated to variations of *JH*).

Tongue model. As the jaw is one of the major tongue carriers, the parameter *JH*, was used as the first control parameter of the tongue model. Its main effect is a tongue rotation around a point in the back of the tongue. The next two parameters, *tongue body TB*, and *tongue dorsum TD*, control respectively the *front-back* and *flattening-bunching* movements of the tongue. The next two parameters, *tongue tip vertical TTV* and *horizontal TTH* parameters, are essentially associated with movements of the tip of the tongue. An extra parameter, related to the hyoid bone height, called *HY*, was used as the sixth control parameter for the tongue model. Altogether, the 3D tongue model is controlled by the six articulatory control parameters. The effects of some of these parameters are demonstrated in Fig. 4 which displays tongue shapes for two extreme values (–2 and +2) of the parameter considered, all other parameters being set to zero. Table 1 displays the variance, relative to the total variance of the full 3D coordinates, explained by each component. It appears that an amount of 87 % is explained by our controlled analysis, which is only 6 % below the optimal result from a raw PCA with the same number of components.



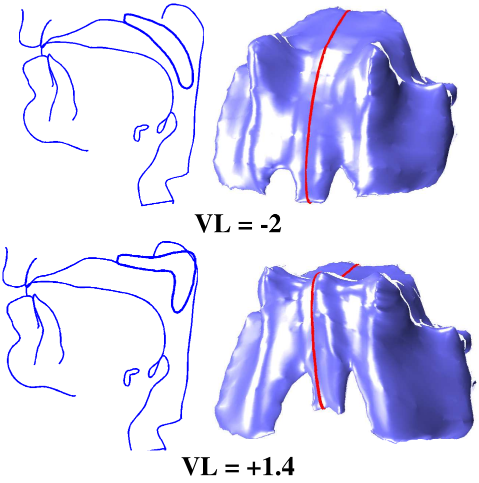
JH = +2

JH = -2

TB = -2

TB = +2

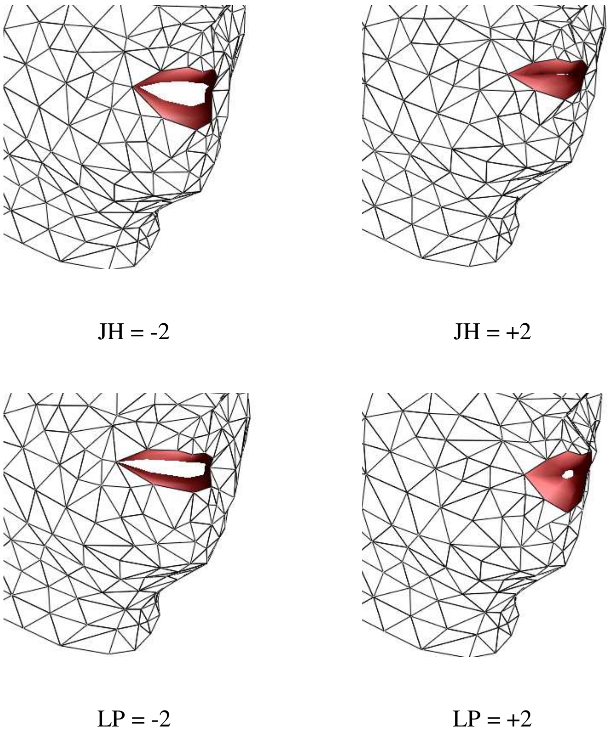
Fig. 4. Jaw and tongue positions for extreme values of JH and TD



VL = -2

VL = +1.4

Fig. 5. Velum shape for two extreme values of VL



JH = -2

JH = +2

LP = -2

LP = +2

Fig. 6. Lips and face shapes for extreme values of JH and LP

Table 1. Relative data variance explained by each component: *varex* is the explained relative data variance, *cum var* its cumulated value

Tongue			Velum			Lips / face		
Param.	varex	cum var	Param.	varex	cum var	Param.	varex	cum var
JH	22,2%	22,2%	VL	83,0%	83,0%	JH	58,5%	58,5%
TB	41,4%	63,6%	VS	6,0%	89,0%	LP	11,0%	69,5%
TD	11,7%	75,3%				LL	16,7%	86,2%
TTV	3,0%	78,4%				UL	3,1%	89,2%
TTH	4,3%	82,6%				JA	2,0%	91,2%
HY	4,5%	87,1%				LY	0,7%	91,9%

Velum. The velum is controlled by two parameters: VL and VS. The effect of parameter VL on the whole velum is illustrated on Fig. 5 by the shape associated with the two extreme values of VL found in the data. The main movement associated with VL is a movement in an oblique direction. Parameter VS is much related to a horizontal displacement coupled with a vertical elongation of the velum, which complements the velopharyngeal port closure by a front to back movement and may significantly modify the velopharyngeal port constriction. These two components explain almost 90 % of the total variance of the data (cf. Table 1 for complete information).

Face and lips model. The face and lips model is controlled by six parameters for neutral speech (cf. [15], or [12] for an extension to more expressive speech). The first one, *JH*, common with the tongue model, controls the opening / closing movement of the jaw and its large influence on lips and face shape (see Fig. 6 for an illustration). Three other parameters are essential for the lips: *LP* controls the protrusion / spreading movement common to both lips that characterises the /i/ vs. /y/ opposition; *UL* controls the upper lip raising / lowering movement, useful to realise the labio-dental consonant /f/ for instance; *LL* controls the lower lip lowering / raising movement found in consonant /ʃ/ for which both lips are maximally open while jaw is in a high position. The second jaw parameter, *JA*, is associated with a horizontal forward / backward movement of the jaw that is used in labio-dental articulations such as /f/ for instance. Note finally a parameter *LX* related to a movement of larynx lowering. Altogether, more than 90% of the data variance in the corpus is taken into account by these components (cf. Table 1 for more detailed information).

3 The Audiovisual Talking Head

Our audiovisual talking head is made of the assemblage of the individual articulators three-dimensional surface models described above. By construction, all models are properly aligned with the common reference coordinate system related to the skull. Fig. 7 illustrates various possible displays of this talking head. The face and lips can be textured to achieve video-realist rendering. Any part of any articulator can be cut away to provide direct visual access to other articulators that may be located more internally in the head. Alternatively, the skin of the face can be made semi-transparent to allow the vision of the inner articulators while keeping a visual reference to the ordinarily visible face and lips. Altogether, these characteristics allow the possibility to produce *augmented* audiovisual speech, i.e. audiovisual speech complemented with displays that cannot be possible in the real life.

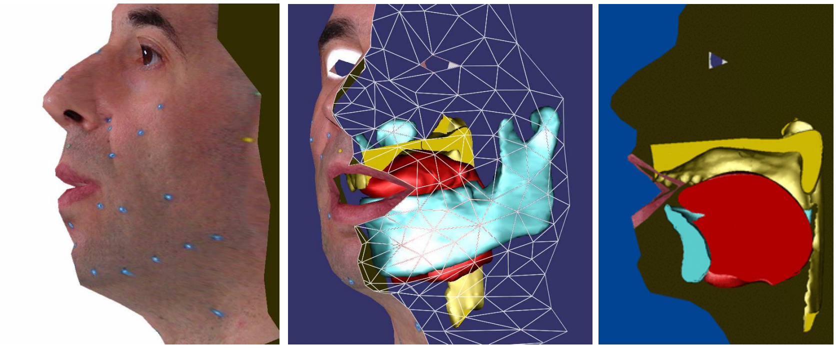


Fig. 7. Examples of displays of the talking head. The face, the jaw, the tongue and the vocal tract walls including the hard and soft palates can be distinguished when needed.

4 Animation of the Talking Head by Motion Capture

4.1 Motion Capture by Means of Electro-magnetic Articulography

Various systems possess some of the motion capture abilities needed to track speech articulators as needed to control and animate a talking head such as that described above: marker-based methods using multiple standard or infrared video cameras to track passive or active light emitting markers attached to the speaker's skin; Electro-Magnetic Articulography (EMA) that allows inferring the coordinates of small electromagnetic receptor coils from the magnetic fields received by these coils from electromagnetic transmitters; medical imaging techniques such as ultrasonic echography, cineradiography or dynamic MRI. However none is perfect: the marker-based techniques allow tracking points only outside the vocal tract; ultrasonic echography is limited mainly to the tongue; cineradiography is too hazardous to allow recording long corpuses; dynamic MRI is still too slow. At present, EMA (cf. [16] or [17]) presents a good compromise: it can track simultaneously up to about 15 points inside and outside the vocal tract, with a typical sampling frequency of 1 kHz, and accuracy better than 0.1 cm; the coils that can be glued to the maxillary, the jaw, the tongue, the velum or the lips (see e.g. Fig. 8) allow tracking *flesh points* i.e. physical locations of the articulators, contrary to medical imaging techniques that provide only contours. A drawback is the poor spatial resolution related to the limited number of points: this can however be dealt with through the high definition articulators models of the talking head, as will be explained in the next section. Finally, note that another important drawback of EMA is its partially invasive nature: the receiver coils have a diameter of about 0.3 cm and must be connected to the device by thin wires that can interfere slightly with the articulation (see illustration in Fig. 8).

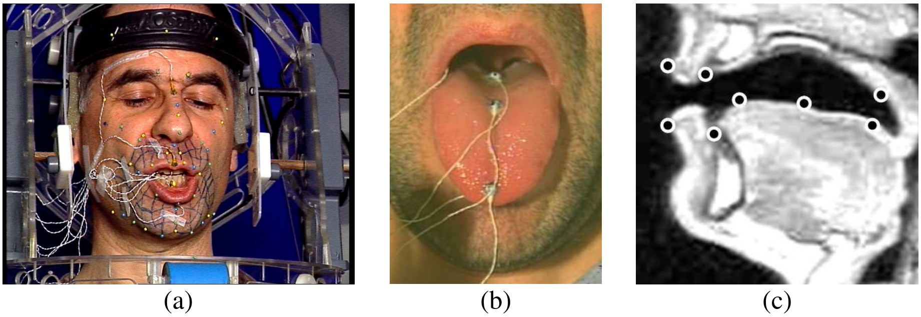


Fig. 8. Illustration of the EMA device: (a) photo of the subject in the EMA device; (b) photo of EMA receptor coils attached to the subject's tongue; (c) possible locations of eight coils (white disks with black centers) in the midsagittal plane

4.2 Control of the Talking Head by Inversion from EMA Recordings

In this section, we show how to control the talking head by inversion from the EMA recordings made on the speaker used for the development of this talking head. We have shown that the various articulators' models can be controlled by a limited number of control parameters: this means that measuring the coordinates of a limited number of carefully chosen points can provide the information necessary to define the complete 3D shape of these articulators. More precisely, as our articulators' models are linear, the 3D coordinates of each vertex are linear combinations of the control parameters. These coordinates can thus be simply obtained by multiplying the vector of control parameters by the matrix of the models' coefficients. Recovering, by inversion, the control parameters of these models can therefore be done from a sufficient number of independent geometric measurements using the (pseudo-) inverse of the model coefficient matrix.

We have thus used the vertical and horizontal coordinates in the midsagittal plane of a set of six coils of an EMA system: a *jaw coil* was attached to the lower incisors, while a *tip coil*, a *mid coil* and a *back coil* were respectively attached at approximately 1.2 cm, 4.2 cm and 7.3 cm from the tongue extremity; an *upper lip coil* a *lower lip coil* were attached to the boundaries between the vermilion and the skin in the midsagittal plane. Extra coils attached to the upper incisors and to the nose served as references. After appropriate scaling and alignment, the coordinates of the coils were obtained in the same coordinate system as the models. No inter speaker normalisation was necessary, since the same speaker is used both for the models and the EMA measurements.

A specific vertex of the 3D tongue mesh was then chosen and associated to each tongue coil in such a way as to minimise the maximum of the distance between the vertex and the coil for a set of articulations representative of the articulatory space of the speaker. The vertices of the lips / face model surface mesh located at the boundary between the vermilion and the skin for each lip in the midsagittal plane were naturally associated with the lips coils. As a first approximation, the left-right coordinates of all the vertices in the midsagittal plane were assumed to be zero, which turned out to be a valid assumption. The three tongue vertices and the two lip vertices associated with the EMA coils have respectively six and four independent coordinates controlled by five parameters each. These control parameters can finally be obtained from the (pseudo-) inverse matrices of the sub-models that predict the coordinates of these specific vertices.

As the *JH* and *JA* parameters are directly proportional to the vertical and horizontal coordinates of the jaw coil, they were computed first in practice, and their linear contributions were removed from the measured EMA coordinates. Parameter *HY* was set to zero, as it was impossible to recover it accurately. The four remaining tongue parameters (*TB*, *TD*, *TTV*, *TTH*) were next obtained by the pseudo-inverse of the matrix for the three tongue coils, and the three remaining lips / face parameters were obtained by the pseudo-inverse of the matrix for the two lips coils. The pseudo-inverse matrices give only sub-optimal solution to the inversion problem, since the number of control parameters is lower than the number of measured coordinates. The mean estimation error, defined as the RMS of the distance between the EMA coils and their associated tongue vertices over a set of 44 articulations was 0.17 cm.

Although in most cases this inversion procedure yields satisfactory results, the resulting tongue contours sometimes cross the hard palate contours, as a consequence of the not modelled nonlinear effect of tongue tip compression when in contact with the palate. In such case, the four tongue parameters are slightly adjusted, using a constrained optimisation procedure that minimises the distance between the coils and the three specific tongue model vertices, with the constraint of preventing the tongue contour from crossing the palate contour.

The video clip PB_phrm6.avi contains an animation obtained by inversion for a sentence in French.

Note finally, that for practical reasons, no coil was attached to the velum in this experiment, and thus no nasal sounds were involved in the study, though we showed in another study that velum could be accurately reconstructed from one EMA coil attached about half way between the hard palate-velum junction and the tip of the uvula ([13]).

5 Example of Application and Perspectives

We have shown that we can control and animate the talking head in a very natural manner from dynamic measurement on a real subject, following a classical paradigm of motion capture, based on EMA recordings. Using this method, we have made a first audiovisual perception experiment aiming at assessing the human tongue reading abilities ([18]), that we briefly present below. We have attempted to determine if direct and full vision of the tongue can be used, based on the augmented speech capabilities of our talking head (in this study, a cutaway profile view). Using the motion capture paradigm based on EMA, we have elaborated a set of audiovisual Vowel-Consonant-Vowel stimuli. These stimuli have been presented to a group of listeners in a series of audiovisual perception experiments using various presentation conditions (audio signal + cutaway view along the sagittal plane *without* tongue, audio signal + cutaway view *with* tongue, audio signal + complete face with skin texture). Each condition was played at four different Signal to Noise Ratios (SNRs) of white noise added to the sound. The analysis of the results has shown some implicit learning effects of *tongue reading*, a preference for a more ecological rendering of the complete face compared to a cutaway presentation, a predominance of lip reading over tongue reading (except for cases where – the audio signal being so much degraded or absent – tongue reading is taking over). These preliminary results need to be complemented by more systematic tests implying notably visual attention measurements.

The talking head appears to possess flexible and various capabilities of augmented audiovisual speech generation: this points the way to a number of promising applications in the domain of speech therapy for speech retarded children, perception and production rehabilitation of hearing impaired children, and pronunciation training for second language learners.

References

1. Mills, A.E.: The development of phonology in the blind child. In: Dodd, B., Campbell, R. (eds.) Hearing by eye: the psychology of lipreading, pp. 145–161. Lawrence Erlbaum Associates, London (1987)
2. Vihman, M.M., Macken, M.A., Miller, R., Simmons, H., Miller, J.: From babbling to speech: A re-assessment of the continuity issue, Language, vol. 61, pp. 397–445 (1985)

3. Stoel-Gammon, C.: Prelinguistic vocalizations of Hearing-Impaired and Normally Hearing subjects. A comparison of consonantal inventories. Journal of Speech and Hearing Disorders 53, 302–315 (1988)
4. Mulford, R.: First words of the blind child. In: Smith, M.D., Locke, J.L. (eds.) The emergent lexicon: The child's development of a linguistic vocabulary, pp. 293–338. Academic Press, New-York (1988)
5. Sumby, W.H., Pollack, I.: Visual contribution to speech intelligibility in noise. Journal of the Acoustical Society of America 26, 212–215 (1954)
6. Benoît, C., Le Goff, B.: Audio-visual speech synthesis from French text: Eight years of models, designs and evaluation at the ICP. Speech Communication 26, 117–129 (1998)
7. Montgomery, D.: Do dyslexics have difficulty accessing articulatory information? Psychological Research 43 (1981)
8. Massaro, D.W., Light, J.: Using visible speech to train perception and production of speech for individuals with hearing loss. Journal of Speech, Language, and Hearing Research 47, 304–320 (2004)
9. Bälter, O., Engwall, O., Öster, A.-M., Kjellström, H.: Wizard-of-Oz Test of ARTUR - a Computer-Based Speech Training System with Articulation Correction. In: Proceedings of the Seventh International ACM SIGACCESS Conference on Computers and Accessibility, Baltimore (2005)
10. Badin, P., Bailly, G., Revéret, L., Baciu, M., Segebarth, C., Savariaux, C.: Three-dimensional linear articulatory modeling of tongue, lips and face, based on MRI and video images. Journal of Phonetics 30, 533–553 (2002)
11. Badin, P., Serrurier, A.: Three-dimensional linear modeling of tongue: Articulatory data and models. In: Proceedings of the 7th International Seminar on Speech Production, ISSP7, Ubatuba, SP, Brazil (2006)
12. Bailly, G., Elisei, F., Badin, P., Savariaux, C.: Degrees of freedom of facial movements in face-to-face conversational speech. In: Proceedings of the International Workshop on Multimodal Corpora., Genoa, Italy (2006)
13. Serrurier, A., Badin, P.: A three-dimensional articulatory model of nasals based on MRI and CT data. Journal of the Acoustical Society of America 123, 2335–2355 (2008)
14. Kelso, J.A.S., Saltzman, E.L., Tuller, B.: The dynamical theory of speech production: Data and theory. Journal of Phonetics 14, 29–60 (1986)
15. Bailly, G., Bérar, M., Elisei, F., Odisio, M.: Audiovisual speech synthesis. International Journal of Speech Technology 6, 331–346 (2003)
16. Perkell, J.S., Cohen, M.M., Svirsky, M.A., Matthies, M.L., Garabieta, I., Jackson, M.T.T.: Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. Journal of the Acoustical Society of America 92, 3078–3096 (1992)
17. Hoole, P., Nguyen, N.: Electromagnetic Articulography in coarticulation research. Forschungsberichte des Instituts für Phonetik und Spachliche Kommunikation der Universität München, vol. 35, pp. 177–184. FIPKM (1997)
18. Tarabalka, Y., Badin, P., Elisei, F., Bailly, G.: Can you read tongue movements? Evaluation of the contribution of tongue display to speech understanding. In: 1ère Conférence internationale sur l'accessibilité et les systèmes de suppléance aux personnes en situation de handicaps (ASSISTH 2007), Toulouse, France (2007)