Gibert et al. Computational Cognitive Science (2015) 1:7
DOI 10.1186/s40469-015-0007-8

Computational Cognitive Science
a SpringerOpen Journal

RESEARCH ARTICLE Open Access

Transforming an embodied conversational agent into an efficient talking head: from keyframe-based animation to multimodal concatenation synthesis

Guillaume Gibert[1,2,3,4*], Kirk N. Olsen[1], Yvonne Leung[1] and Catherine J. Stevens[1]

Abstract

Background: Virtual humans have become part of our everyday life (movies, internet, and computer games). Even though they are becoming 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.

Methods: We describe a method to convert a virtual human avatar (animated through key frames and interpolation) into a more naturalistic talking head. In fact, speech articulation cannot be accurately replicated using interpolation between key frames and talking heads with good speech capabilities are derived from real speech production data. Motion capture data are commonly used to provide accurate facial motion for visible speech articulators (jaw and lips) synchronous with acoustics. To access tongue trajectories (partially occluded speech articulator), electromagnetic articulography (EMA) is often used. We recorded a large database of phonetically-balanced English sentences with synchronous EMA, motion capture data, and acoustics. An articulatory model was computed on this database to recover missing data and to provide 'normalized' animation (i.e., articulatory) parameters. In addition, semi-automatic segmentation was performed on the acoustic stream. A dictionary of multimodal Australian English diphones was created. It is composed of the variation of the articulatory parameters between all the successive stable allophones.

Results: The avatar's facial key frames were converted into articulatory parameters steering its speech articulators (jaw, lips and tongue). The speech production database was used to drive the Embodied Conversational Agent (ECA) and to enhance its speech capabilities. A Text-To-Auditory Visual Speech synthesizer was created based on the MaryTTS software and on the diphone dictionary derived from the speech production database.

Conclusions: We describe a method to transform an ECA with generic tongue model and animation by key frames into a talking head that displays naturalistic tongue, jaw and lip motions. Thanks to a multimodal speech production database, a Text-To-Auditory Visual Speech synthesizer drives the ECA's facial movements enhancing its speech capabilities.

Keywords: Embodied conversational agent; Facial animation; Talking head; Motion capture; Multimodal speech synthesis

* Correspondence: guillaume.gibert@inserm.fr
[1]The MARCS Institute, University of Western Sydney, Locked Bag 1797, Penrith, NSW 2751, Australia
[2]INSERM U846, 18 avenue Doyen Lépine, 69500 Bron, France
Full list of author information is available at the end of the article

Background

Embodied Conversational Agents (ECAs) can use verbal and nonverbal channels of communication to interact with human partners. On the one hand and in the ECA research community, a large amount of work has been devoted to improve ECAs' capabilities to communicate by implementing human-like facial expressions, body gesture, and body posture (Pelachaud 2009), but relatively little research has focused on speech capacities in ECAs. Therefore, ECAs have strong dialog capabilities but weak speech production capabilities (Gris et al. 2014). In general, ECA animation is driven using MPEG-4 Facial Animation Parameters (FAPs). Unfortunately, most FAPs are low-level parameters that do not take into account speech specific gestures (Bailly et al. 2003). On the other hand, the speech research community has focused on visual speech capacities by developing dedicated virtual agents called talking heads. Early talking heads were based on simple animation techniques using a set of key frames (most of the time, visemes (Fisher 1968); i.e., a visually distinguishable face/mouth shape unit) coupled with a set of rules to create the correct transitions between those key frames (Cohen and Massaro 1993). However, this approach is not sufficient to create high quality auditory-visual (AV) speech synthesis. Research has now progressed with new animation techniques using corpora of multimodal speech uttered by humans. The corpora may be based on videos, such as those used in (Ezzat and Poggio 2000; Cosatto and Graf 2000). These approaches have created high quality transitions between visemes that have resulted in impressive synthetic AV speech. In fact, naive subjects are not able to distinguish real from synthetic stimuli (Turing test) (Ezzat et al. 2002). Nevertheless, these systems fail to increase intelligibility when compared to audio-only stimuli (Geiger et al. 2003). Furthermore, some synthesizers need 3D data in order to capture speech articulation. Photogrammetric recordings with beans glued to a speaker's face can provide high resolution facial movement data (Bailly et al. 2002). This passive sensor approach necessitates a tedious preprocessing phase to construct a model that is able to fit unseen data. Other recording techniques have used active sensors to retrieve the positions of sensors over time without necessitating preprocessing. Kuratate (Kuratate 2008) used an Optotrak device to record a speaker's facial movement while uttering speech and created a high quality system able to synthesize AV speech from any text input. Other active sensor equipment has been used such as Electro-Magnetic Articulograph (EMA) to record three dimensional speech articulation database to drive data-driven talking heads (Sheng et al. 2011) for computer assisted pronunciation training.

Humans commonly employ speech reading in adverse listening conditions to facilitate speech perception (Sumby and Pollack 1954). The ability to visually obtain phonetic information depends on watching facial movements that are produced by the speech articulators: mainly by the lips and jaw, and to some extent the larynx and tongue. These movements have been shown to be highly correlated with speech acoustics (Yehia et al. 2002). Although the tongue is a partly occluded speech articulator, its movements provide useful information for visual speech perception. Perceivers perform better with point-light displays including additional dots on the tongue and the teeth than with displays with 'lips only' dots during speech perception experiments (Rosenblum et al. 1996). Accurate 3D tongue models have been included in talking heads. These models are usually obtained by Magnetic Resonance Imaging (Badin et al. 2002; Engwall 2000; Badin et al. 2008) and animated by electromagnetic articulography data (Engwall 2003; Gibert et al. 2012; Steiner et al. 2013) or ultrasound images (Fabre et al. 2014). Computational approaches such as convolutive Nonnegative Matrix Factorization could be used to derive interpretable movement primitives from speech production data (Ramanarayanan et al. 2013). These speech movement primitives can be used to animate virtual agents' speech articulators for a given set of activation data. EMA may be also used to synthesize acoustic speech from the variation of articulatory parameters (Toutios et al. 2013). A promising next step is to use synchronous recordings from EMA and motion capture data systems (Jiang et al. 2002; Engwall 2005). With such a setup, a large number of sensors can be placed on the speaker's face and tongue.

In the present paper, we propose an innovative method to transform an existing ECA animated by interpolation between key frames (i.e., with poor speech capabilities) into a talking head. First, we describe the recording and processing of a multimodal synchronous speech database. An Optotrak Certus (Northern Digital Inc.) motion capture system and a Wave (Northern Digital Inc.) electromagnetic articulography system were used to record an Australian speaker uttering a large set of phonetically-balanced sentences. This unique setup enabled lip, jaw, and tongue trajectories to be recorded synchronously. We built an articulatory model by decomposing each speech articulator movements separately using guided Principal Component Analysis (gPCA). Sensor positions were converted into values of articulatory parameters in order to be used to control most ECAs. Second, the ECA's original animation was modified: face and tongue key frames were transformed into articulatory parameters. Finally, we used the multimodal database to animate the ECA and used the MaryTTS software (Schröder et al. 2011) to create a multimodal text-to-speech synthesizer. This innovative approach can be applied to most ECAs (whose animation module is open) to improve their speech capabilities.

Multimodal database

Method

Setup

An EMA system (Wave, Northern Digital Inc.) and an active motion capture (mocap) system (Optotrak Certus, Northern Digital Inc.) were used to record the position of sensors attached to the face and the tongue during a speech production session. These two systems were manufactured to record synchronously the position of their respective sensors together with the acoustic signal. There were 30 mocap active sensors attached to the speaker's face: 3 for jaw motion, 8 for lip motion, 6 for eyebrow motion and 4 for rigid head motion mounted on a headset. Four additional sensors were attached to the Wave transmitter to align the Optotrak and the Wave referential systems. There were also 6 EMA sensors: 3 glued (using dental glue) on the tongue (tongue tip – **TT**, tongue body – **TB**, and tongue dorsum **TD**) and 1 attached to the nasion and 2 to the tragus. The positions of the sensors on the speaker's face are displayed in Fig. 1.

The EMA field transmitter emits an electromagnetic field and signals transduced in small sensors 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) measurements per sample and per sensor at 100 Hz. The accuracy of the tracking system has been previously assessed and validated for speech research (Berry 2011).

In the study, the positions of 34 Optotrak sensors were recorded at 60 Hz. The positions of the 6 Wave sensors were recorded at 100 Hz. The frames were time-stamped with respect to the beginning of the Optotrak recordings. The audio signal (mono, 22.05 kHz, 16 bits) was recorded synchronously by the EMA system.

Participant

A 30-year old male native speaker of Australian English participated in this recording.

Design and procedure

The speaker was seated close to the Wave transmitter and facing the Optotrak device. The phonetically-balanced sentences of the Lips challenge (Theobald et al. 2008; Theobald 2003) were pronounced by an experimenter and displayed on a screen facing the speaker. The speaker was instructed to repeat each sentence in a neutral tone after the experimenter. A set of 278 phonetically-balanced sentences was recorded. For each sentence a recording session of 10 s was set. Therefore, the total number of frames (at 60 Hz) was 166800.

Data modeling

The recorded data are multimodal in essence: face, tongue trajectories and acoustic signal. Several processing steps were necessary before using the database to animate the ECA. In fact, sensor trajectories cannot be used directly to animate the ECA because of affordance issues. The shape model of an ECA may significantly differ from human morphology. The articulatory modelling provided 'normalized' animation parameters that can be used to control the ECA even if the shape model is different from the speaker's morphology. These processes will be explained in the following subsections.

Acoustic segmentation

The audio files were automatically segmented using the method exposed in the MaryTTS (http://mary.dfki.de/) import voice procedure (Pammi et al. 2010). EHMM acoustic labeler from festvox (Black and Lenzo 2007) was used to generate label files from the audio files and corresponding transcriptions. The label files created by this

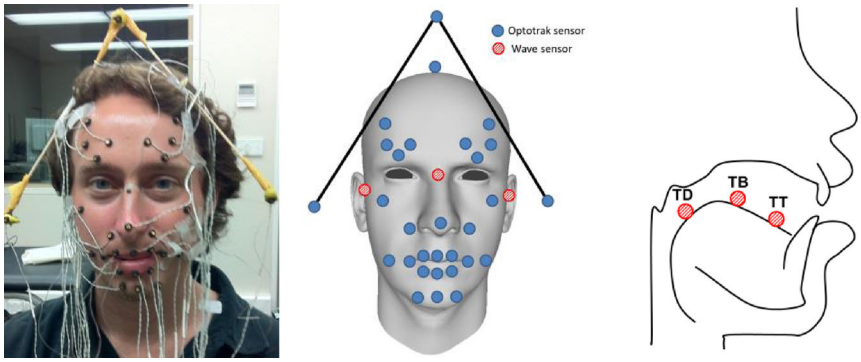

Fig. 1 Sensor positions on the speaker's face and tongue. A headset with 4 sensors was used to estimate the rigid head motion. The Wave sensors (in red) were glued to the tongue tip, tongue body and tongue dorsum and attached to the nasion and tragus. The Optotrak sensors (in blue) were attached to the face and the headset

procedure used the SAMPA phonetic alphabet and were stored as lab files. These files were then converted into Textgrid files by a custom-made Matlab (MathWorks, Inc., Natick, Massachusetts, United States) program. The software Praat (Boersma and Weenink 2010) was used to manually check and correct the segmentation. Therefore, for each audio file, a corresponding text file listed the series of phonemes and their timing information. From these text files, other segmentation files were created containing the series of diphones (i.e., the part of speech comprised between successive stable allophones) and their timing information. These files were used to build the multimodal dictionary that was then used to generate synthetic multimodal speech.

Articulatory model

The Optotrak and Wave devices recorded at different sampling rates. The Wave data were time-stamped by the Optotrak device with respect to the beginning of the Optotrak recording. The Wave data were downsampled (low-pass filtered at 20 Hz) to 60 Hz and, if acquisition timing differed between the Wave and the Optotrak devices, interpolated between two consecutive frames. After this step, the Wave and Optotrak data were sampled at 60 Hz and synchronized.

Head movement (translations and rotations) was estimated and corrected using the Optotrak sensors positioned on the headset and the Wave sensors placed on the nasion and the tragus. Speech articulator movements did not affect these sensors. A modeling procedure was applied to the data with two specific aims: first, to extract meaningful parameters controlling an elementary articulator, and second, to remove artifacts and measure noise. PCA was applied to rigid motion. The first two components explained more than 90 % of the total variance. An articulatory model was built using the method proposed by (Gibert et al. 2005; Revéret et al. 2000; Badin et al. 2002). A pruning step (simple vector quantization) was applied to remove the frames in which sensor positions were too similar (Euclidian distance < 1.0 mm). This step conditioned the data before building the statistical models. Then, the contribution of the different speech articulators (jaw, tongue and lips) and the eyebrows was iteratively subtracted. This subtraction consisted of an iterative application of PCA on subsets of landmarks.

The procedure extracted 10 articulatory parameters; extreme variations of jaw1 and tongue1 are shown in Figs. 2 and 3 respectively:

- Jaw opening (**jaw1**) using PCA on the jaw position sensor values (13.40 % of the global variance);
- Tongue front-back movement (**tongue1**) using PCA on the residual tongue (*TB, TD*) position values (13.17 % of the global variance);
- Tongue flattening-bunching movement (**tongue2**) using PCA on the residual tongue (*TB, TD*) position values (4.83 % of the global variance);

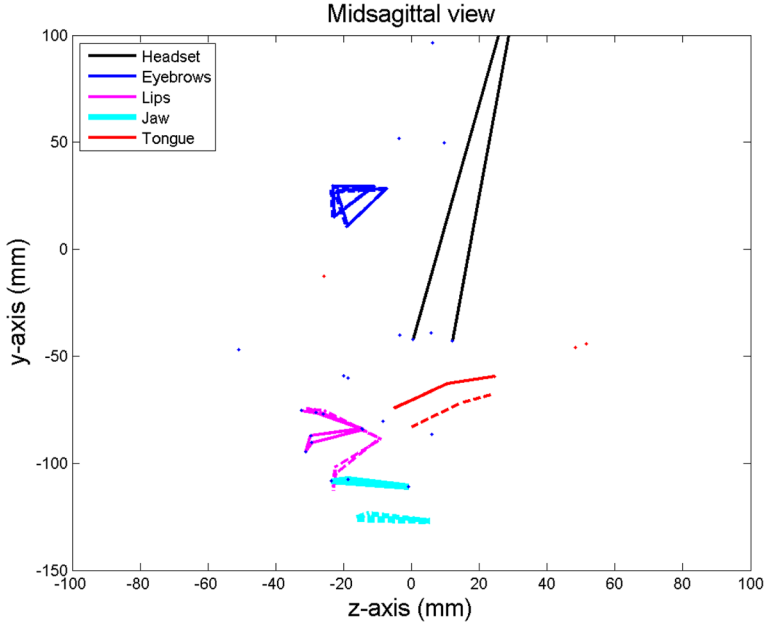

Fig. 2 Maximum variations (solid and dashed lines) of the first articulatory parameter jaw1 driving the jaw. It displays the peaks of opening-closing movement. The tongue follows the jaw opening movement and this movement is encompassed by this articulatory parameter

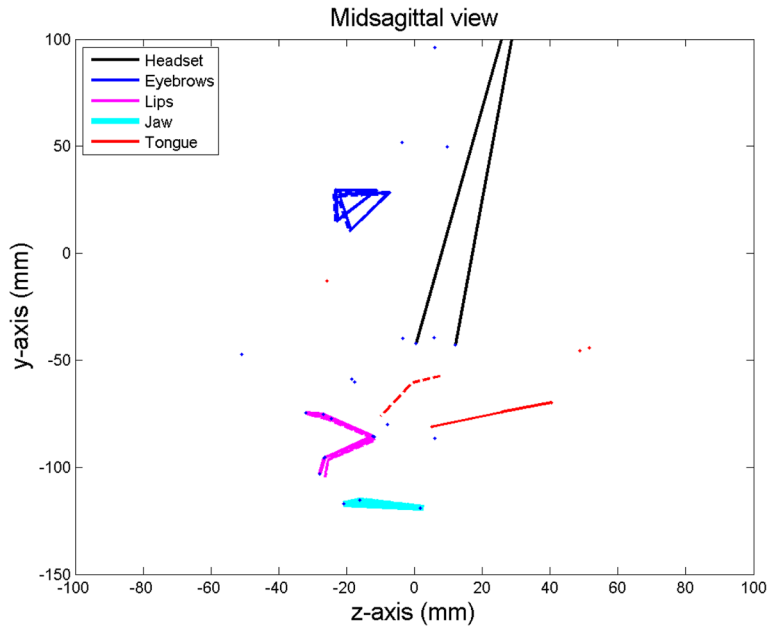

Fig. 3 Maximum variations (solid and dashed lines) of the first articulatory parameter driving the tongue. The parameter tongue1 corresponds to the posterior and anterior extents of a tongue front-back movement

- Tongue tip vertical movement (**tongue3**) using PCA on the residual tongue (*TT*) position values (5.13 % of the global variance);
- Tongue tip horizontal movement (**tongue4**) using PCA on the residual tongue (*TT*) position values (5.69 % of the global variance);
- Lip rounding (**lips1**) using PCA on the residual lip position values (10.06 % of the global variance);
- Lip closing (**lips2**) using PCA on the residual lower lip position values (0.93 % of the global variance);
- Lip raising (**lips3**) using PCA on the residual upper lip position values (2.66 % of the global variance);
- Jaw rotation (**jaw2**) using PCA on the residual jaw position sensor values (2.04 % of the global variance);
- Eyebrow movements (**eyebrows1**) using PCA on the residual eyebrow position values (4.23 % of the global variance).

The maximum variations of the first articulatory parameter driving the jaw (**jaw1**) opening-closing movement are shown in Fig. 2. The tongue is carried by the jaw and this articulatory parameter also drives a tongue rotation around a point at the back of the tongue (Badin and Serrurier 2006). Four additional parameters driving the tongue movements were derived from the data. Figure 3 represents the maximum variations of the first of them. The parameters **tongue1** and **tongue2** were extracted using the position of TB and TD sensors only. They controlled the front-back and flattening-bunching movements. The parameter **tongue3** was extracted by guided PCA using the position of TT sensor only.

Three articulatory parameters driving the lips were extracted. They corresponded to the lip protrusion (**lips1**), lip closing (**lip2**) and lip raising (**lip3**) movements. Finally, a parameter (**eyebrows1**) driving the eyebrow movements was extracted as it may convey prosodic information (Granstrom and House 2005).

Data reconstruction

The rigid head motion was estimated in terms of translations (Tx, Ty, Tz) and rotations (Rx, Ry, Rz) around the neck (the center of rotation was estimated at the same time). These movements were transformed into articulatory parameters using the following equations:

$$\hat{\alpha_H} = \underset{\alpha_H \in [-3;3]^N}{\operatorname{argmin}} \| \hat{P3D}_{HEADSET} - P3D_{HEADSET} \|_2$$

$$\widehat{P3D}_{HEADSET} = RigidMotion\left(m_{HEADSET},\ m_{Hmvt} + \alpha_H\ eigv_H\right)$$

where $\boldsymbol{P3D_{HEADSET}}$ corresponded to the actual position of Optotrak sensors placed on the headset, and $\boldsymbol{\widehat{P3D}_{HEADSET}}$ corresponded to the estimated ones, $\boldsymbol{m_{HEADSET}}$ corresponded to the average position of the sensors on the headset, $\boldsymbol{m_{Hmvt}}$ corresponded to the mean rigid head motion of the model, $\boldsymbol{\alpha_H}$ corresponded to the rigid motion parameter values to be estimated and $\boldsymbol{eigv_H}$ corresponded to the rigid motion model derived from PCA. The number

N of rigid motion parameters driving the headset was 6. The values of each parameter were limited to [−3; 3].

Each recording was then 'inversed', i.e., for each frame of the recording, the values of the articulatory parameters were estimated to minimize the Euclidian distance between the original data and the reconstructed ones using the following equation:

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

where $P3D_{WAVEOPTO}$ corresponded to the position of the Wave and Optotrak sensors after subtraction of the rigid head motion, m_{Face} corresponded to the mean face configuration of the articulatory model, α corresponded to the articulatory parameter values and $eigv_{Face}$ corresponded to the articulatory model. The number M of articulatory parameters driving the tongue was 10. The values of each parameter were limited to [−3; 3]. This inversion used the relation between the relative sensor positions to recover the position of missing data. Therefore, the sensor trajectories were transformed into variations of articulatory parameter values.

A low-pass Butterworth filter (6^{th} order, 8Hz) was applied to the articulatory parameters in order to remove noise due to recordings and missing data. An example of variation of the articulatory parameter values across time for a sentence of the corpus can be seen in Fig. 4. The variation of these parameters is clearly nonlinear. Animation approaches that use linear interpolation between key frames cannot replicate such variations. After this filtering, the recordings were reconstructed, i.e., missing data were estimated using the articulatory model. The reconstruction error computed as the Euclidian distance between the recorded position of the Optotrak/Wave sensors and the reconstructed ones was M = 5.41 mm and SD = 3.15 mm. An example of data reconstruction can be found in the Additional file 1: Video S1. This dataset comprised a large amount of English phonemes in different contexts. A multimodal diphone dictionary was created. It contained for each diphone, the variations of the articulatory parameters driving the visible and partly occluded speech effectors.

Embodied Conversational Agent

Avatar

The avatar used in this study was a representation of an 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

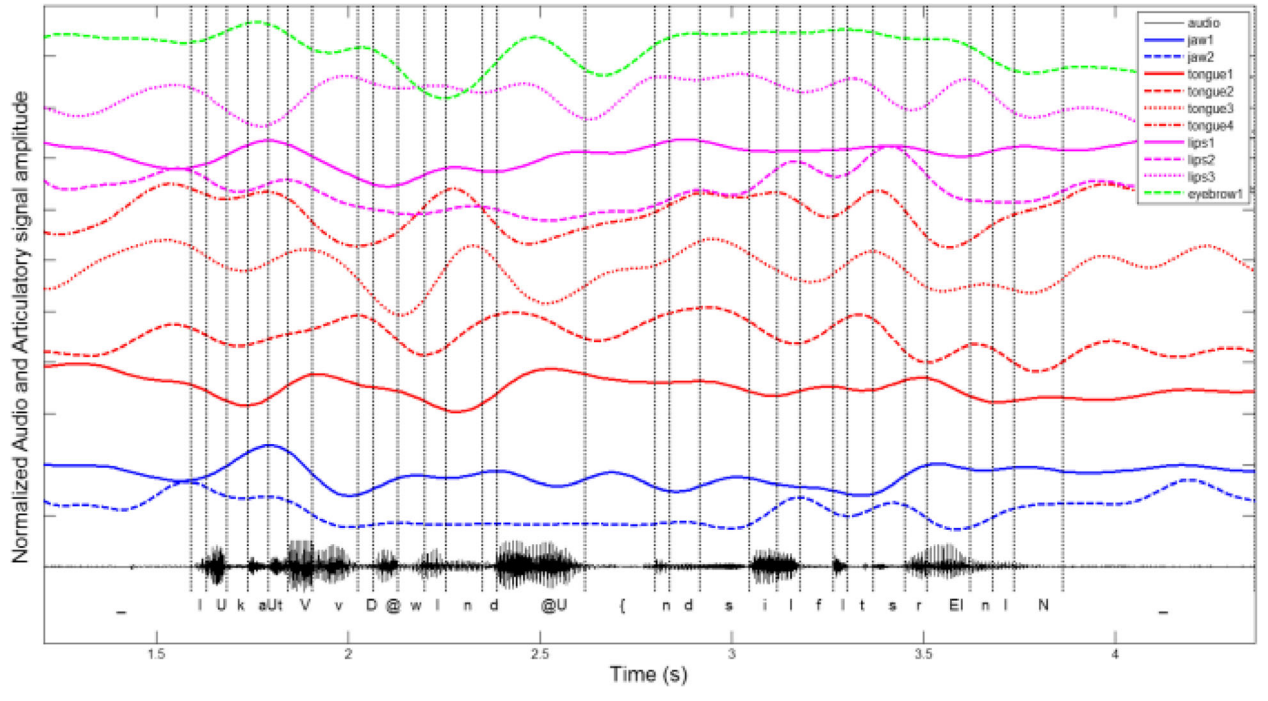

Fig. 4 Variation of the ten articulatory parameter values for the sentence: "Look at the windows and see if it's raining". The variations are nonlinear and cannot be replicated with linear interpolation animation. The database is segmented into phonemes. A multimodal diphone dictionary is created with this segmented database

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.

Face articulatory parameters

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. This parameter recovered 2 % of the global variance in the EMA data. It was not included in the final set of synthetic articulatory parameters for the animation.

Tongue articulatory parameters

Because tongue key frames were not related to any speech articulation in the original animation, but only to meaningless geometric variations, an alternative method was designed. Each tongue sensor from EMA data was associated with a specific vertex of the 3D tongue mesh (which is composed of 50 vertices) of the original ECA 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 $\boldsymbol{P3D_{K_i}}$ corresponded to the position of the three selected vertices of the 3D tongue mesh for the key frame $\boldsymbol{K_i}$, $\boldsymbol{\alpha_i}$ corresponded to the weights applied to the key frame $\boldsymbol{K_i}$, and $\boldsymbol{P3D_{EMA}}$ 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 $\boldsymbol{\alpha_i}$ were limited to [−10; 10]. Examples of configurations found in the EMA database and the corresponding constrained 3D tongue mesh are visualized in Fig. 5.

After this step, a quantized database of 3D tongue postures was created. For all the configurations of the EMA database, corresponding constrained 3D tongue mesh of 50 vertices 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 4.2.2 was used to build a tongue articulatory model using the database of 3D tongue postures in addition to the EMA database. Finally, the articulatory tongue model was controlled by 5 articulatory parameters (as described in (Badin and Serrurier 2006)): 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 Fig. 6. This avatar was driven by similar articulatory parameters to the ones derived in the modelling procedure described above (Gibert et al. 2012).

Auditory-Visual Text-To-Speech system

Overview

The system proceeds in three steps. First, a module connects to the MaryTTS server (http://mary.dfki.de/) and asks for the list of phonemes and its duration for a given text. Second, the same module requests an acoustic signal corresponding to the given text. Third, the list of phonemes plus duration is sent to the Visual Text-To-Speech (VTTS) module, which searches the best list of diphone candidates and concatenates them to create articulatory parameter trajectories. The acoustic signal and the articulatory parameter trajectories are sent to the animation module, which plays the data i.e., the ECA speaks and moves his speech effectors (jaw, lips and tongue) accordingly. The schematic representation of the system is shown in Fig. 7.

Text-to-Speech

The first step of our system is to ask the MaryTTS server (Schröder et al. 2011) to generate an acoustic signal that corresponds to the new sentence. To this end, the MaryTTS server performs natural language processing to generate a list of phonemes and prosodic information (duration, pitch variation, intensity). Finally, the synthesizer creates a sound signal from this information using a cluster unit selection code derived from FreeTTS. In fact, the recorded speech signal was used to create a specific MaryTTS synthetic voice for our speaker following the procedure described in the voice creation module (Pammi et al. 2010). This procedure performs feature extraction from acoustic and text data and then automatic segmentation/labelling. The procedure was bypassed at this stage to check manually the automatic segmentation/labelling (see subsection 4.2.1). Then, the system builds a unit selection voice from the manually

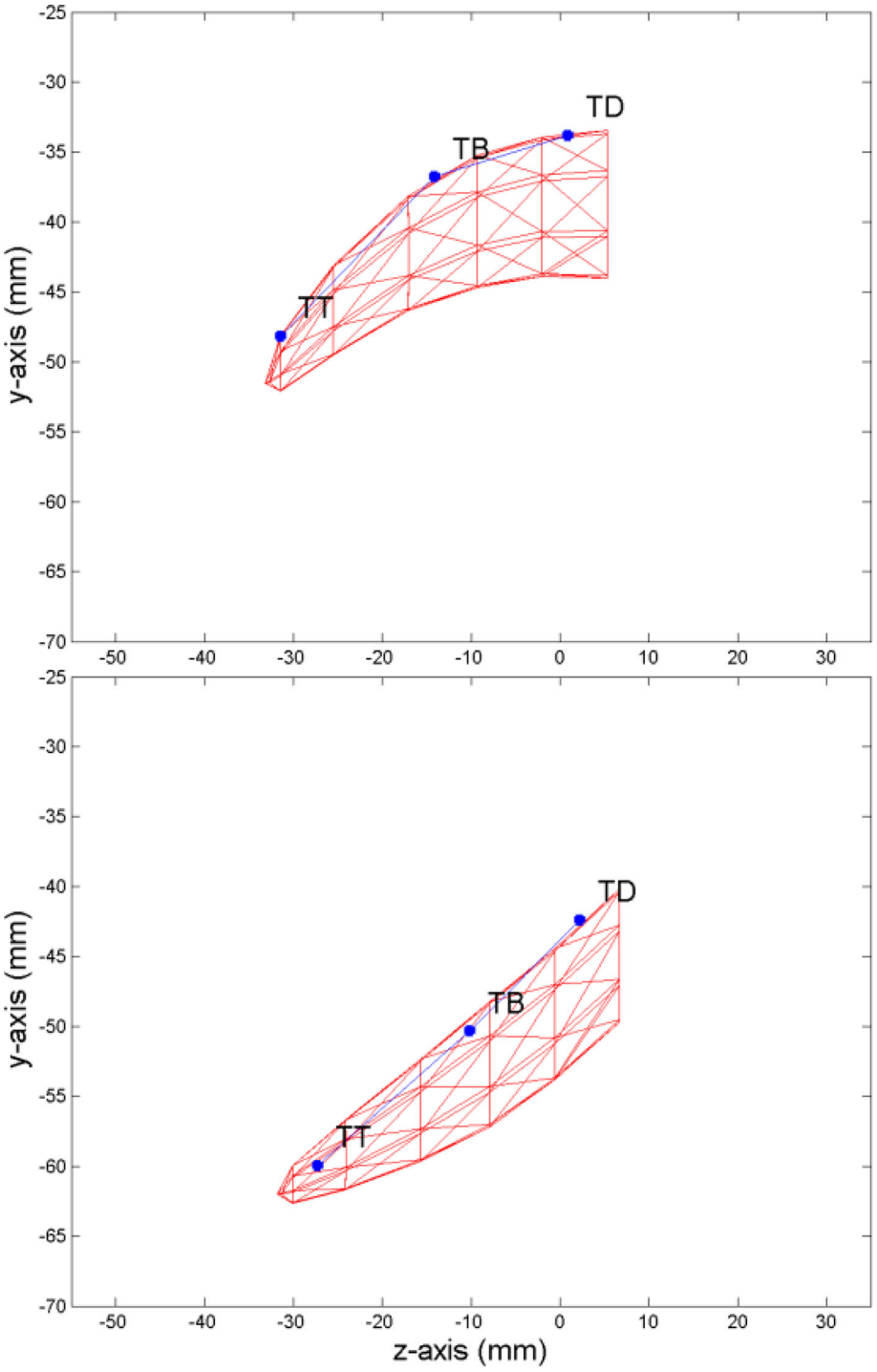

Fig. 5 Two tongue configurations (midsagittal view) from the quantized EMA database (TD, TB and TT sensor positions in blue) and the corresponding constrained 3D tongue mesh (red mesh)

checked segmentation files and the acoustic features. Given the manual segmentation, the quality of the voice is better than the fully automatic procedure that can generate artifacts because of segmentation errors.

The second step consists of asking the MaryTTS server (Schröder et al. 2011) to provide the list of phonemes with their duration. This information is sent to the visual synthesis system. The generation of visual speech is explained in the following section.

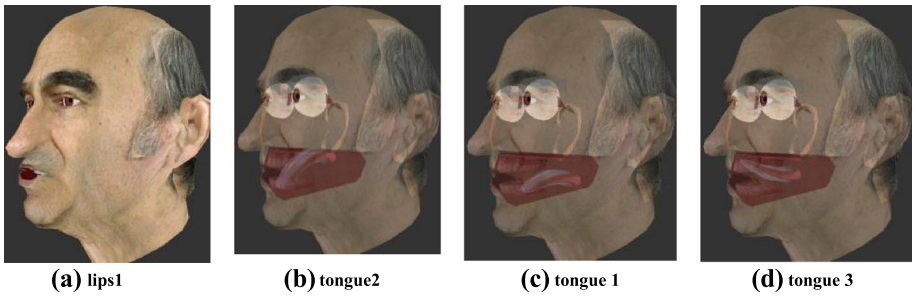

Fig. 6 Examples of the maximum variation (one direction) of some articulatory parameters driving the avatar. **a lips1** corresponds to lip protrusion; **b tongue2** corresponds to the flattening-bunching movement, **c tongue1** corresponds to the posterior extent of a tongue front-back movement, **d tongue3** corresponds to the peak of a tongue tip vertical movement

Visual synthesis

Given the list of phonemes and their duration, the visual synthesis system creates a list of diphones and their corresponding duration. For instance, if the word "Welcome" is submitted to the MaryTTS server, it will return the following list of phones: _ w E l k @ m _ with their respective timing (in seconds): 0.060, 0.125, 0.19, 0.29, 0.385, 0.51, 0.695. The corresponding list of diphones is then derived as follows: _w, wE, El, lk, k@, @m, m_. The system searches in the multimodal diphone dictionary containing the trajectories of the articulatory parameters the various candidates for each diphone. This step generates a trellis of diphones (see Fig. 8). The best series of diphones is then selected using a cost function based on a concatenation weight. This step selects the best series of diphones that minimize the Root Mean Square (RMS) distance between values of the articulatory parameters of the previous diphone and the current diphone.

Except in the case that a series of diphones corresponds to a series contained in the dictionary (for instance, the diphones El, lk and k@ comes from the same sentence in the example of Fig. 8), there is always a gap at each concatenation frontier (for example, between the diphone @m and m_ in the example of Fig. 8). This gap is reduced by applying a gapless processing step on the articulatory parameters within the preceding diphone gradually (Gibert et al. 2005). It consists of adding a small value (equal to Δ* frame_index/frame_total_number, where Δ corresponds to the distance (i.e., gap) between the end of the current diphone and the beginning of next one, frame_index corresponds to the time step and frame_total_number corresponds to the total number of time steps for the current diphone) to the variation of the articulatory parameters at each time step. Even if there is a concatenation gap between two consecutive diphones as shown in Fig. 8, this procedure cancels it while keeping

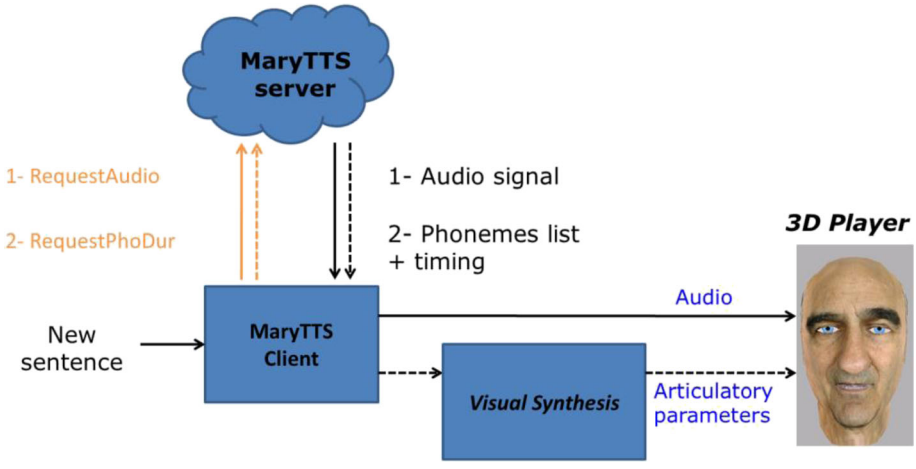

Fig. 7 Schematic representation of the auditory-visual speech synthesis system. Given a new sentence to pronounce, the program acts as a MaryTTS client and asks for an audio signal corresponding to this sentence and to the list of phonemes with their duration (provided by the prosodic module embedded in the MaryTTS software). From the list of phonemes, a second program performs the visual synthesis. It searches the best series of diphones given selection and concatenation costs in the multimodal dictionary. This series is processed to match the expected duration and minimizes the gaps at each boundary. Finally, the acoustic signal and the variation of the articulatory parameters are passed to the 3D Player which animates the ECA accordingly

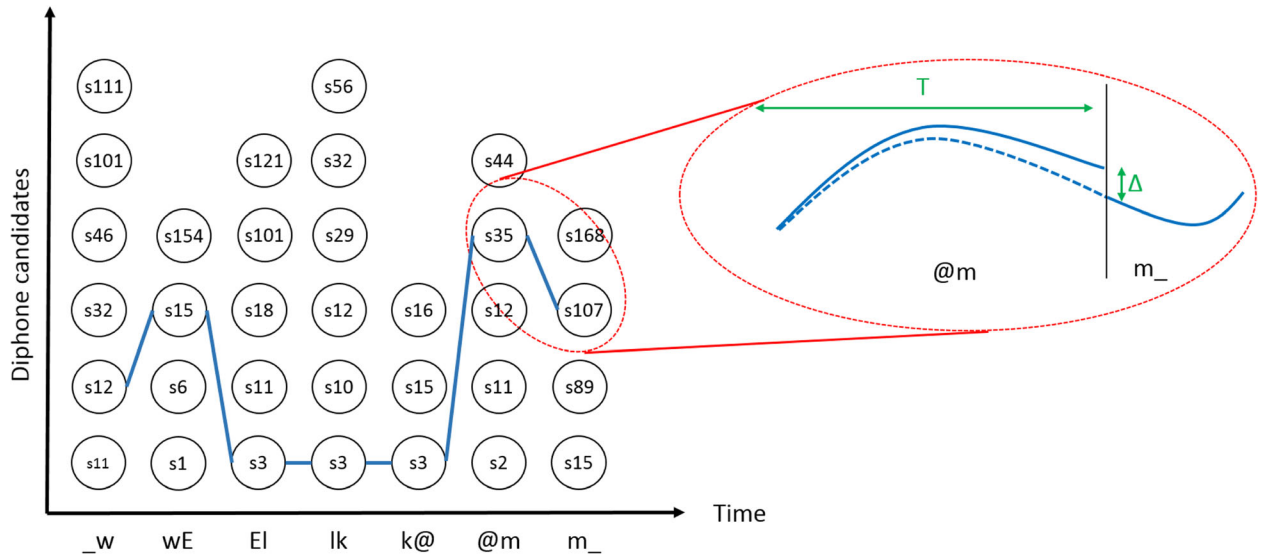

Fig. 8 Schematic representation of the visual synthesis system. Given a series of diphones (e.g., "Welcome"), the system searches in the diphone dictionary the candidates and creates a trellis of diphones (i.e., a list of diphone candidates). The best series of diphones is selected using a concatenation cost, i.e., the aim is to minimize the RMS distance between values of the articulatory parameters of the last frame of the previous diphone and the first frame of the current one. The selected diphones are linked by a blue line in this example. If consecutive diphones come from the same sentence, there is no gap at the concatenation borders (e.g., this is the case for the diphones El, lk and k@ in this example). In other cases, there is a gap (Δ) at the concatenation that must be minimized. A gapless processing step is applied to each articulatory parameter. It consists of adding a small value (equal to Δ/T*frame_index, where Δ corresponds to the difference between the values of the articulatory parameter of the last frame of the current diphone and the first frame of the following one, T corresponds to the total number of frames of the current diphone and frame_index corresponds to the current frame number) at each time step to reduce the gap while keeping the temporal nonlinear variation of the parameter values

the nonlinear variations of the articulatory parameters. This way, the final sample of the previous diphone coincides with the first sample of the current one.

Animation

Acoustic and articulatory data are processed in parallel. The system requests the audio file and the phone list together with timing to the MaryTTS server. The VTTS module creates a file containing the variation of the articulatory parameters with the same timing used by MaryTTS to create the acoustic file (as described in the previous subsection). Once the acoustic signal and the matrix (articulatory parameter x time) of articulatory parameter variations are available, two threads are started: one playing the sound and another one playing the face gestures. The animation module is a custom written software developed using Java and Java3D/JoGL. The animation module plays congruent and synchronous articulatory and acoustic signals from the same Australian speaker. In fact, any existing English MaryTTS voices could be used together with the articulatory data we recorded. However, the visual signals may become incoherent with American or British English voices for instance. Moreover, synchronization may be also different in this case. In our system, the same segmentation files were used to create the acoustic and articulatory sound units. Therefore, this module plays the multimodal data synchronously. An example of animation can be found in the Additional file 1: Video S1.

Even though the articulatory parameters driving the avatar and the ones derived from the speaker have the same topology (e.g., **jaw1** controlled in both cases the jaw opening/closing), it may happen that positive variation of **jaw1** corresponded to jaw opening for the avatar's model and jaw closing for the speaker's model. The sign attribution was determined manually.

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, a multimodal speech database consisting of the recording of face and tongue movements during the production of a large number of sentences by an Australian speaker was created. This database was processed to create a dictionary of synchronous multimodal diphones. An articulatory model of the ECA was then created by transforming selected key frames into articulatory parameters for the jaw, lips and tongue. Real articulatory data together with the acoustic signal were used to steer the talking head using the MaryTTS

software (Schröder et al. 2011) to generate the synthetic acoustic signal and the phoneme and prosodic information. The original ECA with good non-verbal capabilities (facial expressions, blinks, gestures, pupil dilation/constriction (Gibert and Stevens 2012), etc.) and poor speech capabilities was transformed. The ECA kept its nonverbal capabilities intact and significantly improved its speech capabilities. Importantly, this method is a bridge between the ECA and speech communities. More methods should be established to take advantages of the development of the two communities to create virtual agents able to interact naturally with human partners. Human communication is multimodal in essence: verbal, coverbal and nonverbal channels of communication are used during face-to-face communication. Each community has developed specific models for verbal, coverbal and nonverbal behaviors. Bridges such as the proposed method are vitally important for virtual humans to efficiently use all possible channels of communication.

The proposed method could be easily extended to other realistic auditory-visual animation methods based on concatenation (Musti et al. 2011) or Hidden Markov Models (Bailly et al. 2009) using the same unique multimodal database. Electromagnetic articulography provides only a spatially sparse representation of the tongue movements. Co-registration methods of EMA and real-time magnetic resonance imaging (rtMRI) data (Kim et al. 2014) provides richer spatio-temporal data to animate the tongue movements and create a better 3D tongue model. The proposed method could be applied on the USC-TIMIT (Narayanan et al. 2014) which is an extensive database of multimodal (EMA, rtMRI, acoustics) speech production. This work could be extended to emotional speech production which generates different articulation patterns for critical speech articulators compared to neutral speech production (Kim et al. 2015). An evaluation of the method will be performed in the future to assess the gain in intelligibility; for instance, through a speech in noise perception experiment (Gibert et al. 2013). Furthermore, 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 (Gibert et al. 2005) and second language learners (Engwall 2008) to effectively utilize a larger number of virtual agent applications.

Additional file

Additional file 1: Video S1. The video illustrates the articulatory parameters (jaw1, jaw2, tongue1, tongue2, tongue 3, tongue 4, lips1, lips 2, lips3, eyebrows1) derived from the speech production database and their effects on facial movements. Then, an example of data reconstruction is shown, this processing step allows to recover missing data. Finally, an example of ECA animation using real speech production data is played. (MP4 14967 kb)

Competing interests
We have read and understood Springer policy on declaration of interests and declare that we have no competing interests.

Authors' contributions
GG participated in the design of the database. He carried out the statistical modelling to create the articulatory model and adapt it to the Embodied Conversational Agent. He developed the multimodal speech synthesis system. KO participated in the design of the motion capture recording and in the recording of multimodal speech database. He helped to draft the manuscript.YL participated in the recording of the database and in post processing (speech segmentation) of the data. She helped to draft the manuscript. CS participated in the design and coordination and helped to draft the manuscript. All authors read and approved the final manuscript.

Authors' information
GG received a Ph.D. in Signal, Speech and Image Processing in 2006 (Institut National Polytechnique de Grenoble, France) and conducted his doctoral research, entitled "Conception and evaluation of a text-to-Cued Speech 3D system", at the Institut de la Communication Parlée (Grenoble, France). He developed an avatar able to produce Cued Speech (Cued Speech is a manual technique which complements lip-reading for deaf people) from any text input. He set up and ran perception experiments (speech communication methods including also the eye tracking technique) with hearing and deaf people to evaluate the synthesizer. Then, his research interests focused on real-time EEG signal processing and Brain Computer Interfaces (INSERM U821, Lyon, France). After a postdoc at MARCS Institute (UWS, Australia) where he was working in the evaluation Embodied Conversational Agent in the frame of the Thinking Head project, he is now the team leader of the SWoOZ project funded by the ANR PDOC.
KO is an experimental psychologist at the MARCS Institute, University of Western Sydney (UWS). His doctoral dissertation in 2011 investigated mechanisms underlying perceptual and psychophysiological response to acoustic intensity dynamics in speech and music. Dr Olsen has published in the fields of auditory psychophysics, music perception and psychophysiology, and currently holds a UWS career-development post-doctoral research fellowship in the Music Cognition and Action research program at the MARCS Institute.
YL was a research assistant and now a PhD candidate at the MARCS Institute, University of Western Sydney. Her role in this Thinking Head project includes data collection and analyses. Her current PhD project investigates the learning of novel music systems using mathematically generated musical scales.
Cognitive psychologist Catherine J. Stevens adapts experimental methods to i) evaluate complex systems and human-computer interaction; and ii) investigate psychological processes in perceiving and performing music and dance. She holds BA (Hons) and PhD degrees from the University of Sydney and has been an Australian Research Council (ARC) Postdoctoral Fellow at the University of Queensland. She led the Evaluation Team on the interdisciplinary project "From Talking Heads to Thinking Heads". Kate is Professor in the School of Social Sciences and Psychology and is Director of Research and Engagement in the MARCS Institute at the University of Western Sydney (http://marcs.uws.edu.au/).

Acknowledgements
We thank Steve Fazio for his technical support during the recording. This work was supported by the Thinking Head project (Burnham et al. 2006–2011), a special initiative scheme of the Australian Research Council and the National Health and Medical Research Council (TS0669874), by the MARCS Institute, University of Western Sydney and by the SWoOZ project (ANR 11 PDOC 01901).

Author details
[1]The MARCS Institute, University of Western Sydney, Locked Bag 1797, Penrith, NSW 2751, Australia. [2]INSERM U846, 18 avenue Doyen Lépine, 69500 Bron, France. [3]Stem Cell and Brain Research Institute, 69500 Bron, France. [4]Université de Lyon, Université Lyon 1, 69003 Lyon, France.

Received: 22 May 2015 Accepted: 30 August 2015
Published online: 08 September 2015

References

Badin, P., & Serrurier, A. (2006). *Three-dimensional linear modeling of tongue: Articulatory data and models.* Paper presented at the 7th International Seminar on Speech Production, Belo Horizonte, Brazil

Badin, P, Bailly, G, Reveret, L, Baciu, M, Segebarth, C, & Savariaux, C. (2002). Three-dimensional linear articulatory modeling of tongue, lips and face, based on MRI and video images. *Journal of Phonetics, 30*(3), 533–553.

Badin, P., Elisei, F., Bailly, G., & Tarabalka, Y. (2008). An audiovisual talking head for augmented speech generation: Models and animations based on a real speaker's articulatory data. In *Articulated Motion and Deformable Objects, Proceedings* (Vol. 5098, pp. 132–143, Lecture Notes in Computer Science)

Bailly, G., Gibert, G., & Odisio, M (2002). Evaluation of movement generation systems using the point-light technique. In *Speech Synthesis, 2002. Proceedings of 2002 IEEE Workshop on, 2002* (pp. 27–30)

Bailly, G, Berar, M, Elisei, F, & Odisio, M. (2003). Audiovisual Speech Synthesis. *International Journal of Speech Technology, 6*, 331–346.

Bailly, G., Govokhina, O., Elisei, F., & Breton, G. (2009). Lip-synching using speaker-specific articulation, shape and appearance models. *Journal of Acoustics, Speech and Music Processing. Special issue on "Animating Virtual Speakers or Singers from Audio: Lip-Synching Facial Animation"*, doi:10.1155/2009/769494

Berry, JJ. (2011). Accuracy of the NDI Wave Speech Research System. *Journal of Speech, Language, and Hearing Research, 54*(5), 1295–1301. doi:10.1044/1092-4388(2011/10-0226).

Black, A. W., & Lenzo, K. (2007). Festvox: Building synthetic voices. (2.1 ed.)

Boersma, P., & Weenink, D. (2010). Praat: doing phonetics by computer. (5.1.31 ed.)

Burnham, D., Dale, R., Stevens, K., Powers, D., Davis, C., Buchholz, J., et al. (2006–2011). From Talking Heads to Thinking Heads: A Research Platform for Human Communication Science. ARC/NH&MRC Special Initiatives, TS0669874

Cohen, MM, & Massaro, D. (1993). Modeling Coarticulation in Synthetic Visual Speech. In NM Thalmann & D Thalmann (Eds.), *Models and Techniques in Computer Animation*. Tokyo, Japan: Springer.

Cosatto, E, & Graf, H-P. (2000). Photo-realistic talking heads from image samples. *IEEE Transactions on Multimedia, 2*, 152–163.

Engwall, O. (2000). A 3D tongue model based on MRI data. In *International Conference on Spoken Language Processing, Beijing, China* (Vol. 3, pp. 901–904)

Engwall, O. (2003). Combining MRI, EMA and EPG measurements in a three-dimensional tongue model. *Speech Communication, 41*(2–3), 303–329. doi:10.1016/s0167-6393(03)00132-2.

Engwall, O. (2005). Articulatory synthesis using corpus-based estimation of line spectrum pairs. Paper presented at the INTERSPEECH, Lisbon, Portugal.

Engwall, O. (2008). 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)

Ezzat, T, & Poggio, T. (2000). Visual speech synthesis by morphing visemes. *International Journal of Computer Vision, 38*(1), 45–57.

Ezzat, T., Geiger, G., & Poggio, T. (2002). *Trainable videorealistic speech animation.* Paper presented at the ACM SIGGRAPH, San Antonio, TX

Fabre, D., Hueber, T., & Badin, P. (2014). *Automatic animation of an articulatory tongue model from ultrasound images using Gaussian mixture regression.* Paper presented at the INTERSPEECH, Singapore

Fisher, CG. (1968). Confusions Among Visually Perceived Consonants. *Journal of Speech, Language, and Hearing Research, 11*(4), 796–804.

Geiger, G., Ezzat, T., & Poggio, T. (2003). Perceptual Evaluation of Video-realistic Speech. In C. P. #224 (Ed.), *AI Memo #2003-003*. Cambridge, MA: Massachusetts Institute of Technology

Gibert, G., & Stevens, C. J. (2012). *Realistic eye model for Embodied Conversational Agents.* Paper presented at the ACM 3rd International Symposium on Facial Analysis and Animation, Vienna, Austria, 21st September 2012

Gibert, G, Bailly, G, Beautemps, D, Elisei, F, & Brun, R. (2005). Analysis and synthesis of the three-dimensional movements of the head, face, and hand of a speaker using cued speech. *Journal of Acoustical Society of America, 118*(2), 1144–1153. doi:10.1121/1.1944587.

Gibert, G., Attina, V., Tiede, M., Bundgaard-Nielsen, R., Kroos, C., Kasisopa, B., et al. (2012). *Multimodal Speech Animation from Electromagnetic Articulography Data.* Paper presented at the 20th European Signal Processing Conference (EUSIPCO), Bucharest, Romania

Gibert, G, Leung, Y, & Stevens, CJ. (2013). Control of speech-related facial movements of an avatar from video. *Speech Communication, 55*(1), 135–146. http://dx.doi.org/10.1016/j.specom.2012.07.001.

Granstrom, B, & House, D. (2005). Audiovisual representation of prosody in expressive speech communication. *Speech Communication, 46*(3–4), 473–484.

Gris, I, Novick, D, Camacho, A, Rivera, D, Gutierrez, M, & Rayon, A. (2014). Recorded Speech, Virtual Environments, and the Effectiveness of Embodied Conversational Agents. In T Bickmore, S Marsella, & C Sidner (Eds.), *Intelligent Virtual Agents. Vol. 8637, Lecture Notes in Computer Science* (pp. 182–185). New York: Springer International Publishing.

Jiang, J., Alwan, A., Bernstein, L. E., Keating, P., & Auer, E. (2002). *On the correlation between facial movements, tongue movements and speech acoustics.* Paper presented at the International Conference on Spoken Language Processing (ICSLP), Bejing, China

Kim, J, Lammert, AC, Kumar Ghosh, P, & Narayanan, SS. (2014). Co-registration of speech production datasets from electromagnetic articulography and real-time magnetic resonance imaging. *Journal of Acoustical Society of America, 135*(2), EL115–EL121. http://dx.doi.org/10.1121/1.4862880.

Kim, J, Toutios, A, Lee, S, & Narayanan, SS. (2015). A kinematic study of critical and non-critical articulators in emotional speech production. *Journal of Acoustical Society of America, 137*(3), 1411–1429. http://dx.doi.org/10.1121/1.4908284.

Kuratate, T. (2008). *Text-to-AV synthesis system for Thinking Head Project.* Paper presented at the Auditory-Visual Speech Processing, Brisbane, Australia

Musti, U., Toutios, A., Colotte, V., & Ouni, S. (2011). *Introducing Visual Target Cost within an Acoustic-Visual Unit-Selection Speech Synthesizer.* Paper presented at the AVSP, Volterra, Italy

Narayanan, S, Toutios, A, Ramanarayanan, V, Lammert, A, Kim, J, Lee, S, et al. (2014). Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC). *Journal of Acoustical Society of America, 136*(3), 1307–1311. http://dx.doi.org/10.1121/1.4890284.

Pammi, S. C., Charfuelan, M., & Schröder, M. (2010). *DFKI-LT - Multilingual Voice Creation Toolkit for the MARY TTS Platform.* Paper presented at the LREC, Valleta, Malta

Pelachaud, C. (2009). Studies on gesture expressivity for a virtual agent. *Speech Communication, 51*(7), 630–639. doi:10.1016/j.specom.2008.04.009.

Ramanarayanan, V, Goldstein, L, & Narayanan, SS. (2013). Spatio-temporal articulatory movement primitives during speech production: Extraction, interpretation, and validation. *Journal of Acoustical Society of America, 134*(2), 1378–1394. doi:10.1121/1.4812765.

Revéret, L., Bailly, G., & Badin, P. (2000). MOTHER: a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation. In *International Conference on Speech and Language Processing, Beijing, China,* (pp. 755–758)

Rosenblum, LD, Johnson, JA, & Saldana, HM. (1996). Point-light facial displays enhance comprehension of speech in noise. *Journal of Speech and Hearing Research, 39*(6), 1159–1170.

Schröder, M, Charfuelan, M, Pammi, S, & Steiner, I. (2011). Open source voice creation toolkit for the MARY TTS Platform. In *12th Annual Conference of the International Speech Communication Association - Interspeech 2011, Florence, Italy, 2011–08* (pp. 3253–3256). Italy: ISCA. https://hal.inria.fr/hal-00661061/document, https://hal.inria.fr/hal-00661061/file/Interspeech2011.pdf.

Sheng, L., Lan, W., & En, Q. The Phoneme-Level Articulator Dynamics for Pronunciation Animation. In *Asian Language Processing (IALP), 2011 International Conference on, 15–17 Nov. 2011 2011* (pp. 283–286). doi:10.1109/ialp.2011.13

Steiner, I., Richmond, K., & Ouni, S. (2013). *Speech animation using electromagnetic articulography as motion capture data.* Paper presented at the Auditory-Visual Speech Processing (AVSP), Annecy, France, August 29 - September 1, 2013

Sumby, WH, & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of Acoustical Society of America, 26*, 212–215.

Theobald, BJ. (2003). *Visual speech synthesis using shape and appearance models*. Norwich, UK: University of East Anglia.

Theobald, B. J., Fagel, S., Bailly, G., & Elisei, F. (2008). *LIPS 2008: Visual Speech Synthesis Challenge.* Paper presented at the INTERSPEECH 2008, Brisbane, Australia

Toutios, A., Shrikanth, S., & Narayanan, S. (2013). *Articulatory Synthesis of French Connected Speech from EMA Data.* Paper presented at the INTERSPEECH, Lyon, France

Yehia, HC, Kuratate, T, & Vatikiotis-Bateson, E. (2002). Linking facial animation, head motion and speech acoustics. *Journal of Phonetics, 30*(3), 555–568.