Control of speech-related facial movements of an avatar from video

Guillaume Gibert[1,2,3,4], Yvonne Leung[4] & Catherine J. Stevens[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]School of Social Sciences & Psychology, University of Western Sydney, Locked Bag 1797, Penrith, NSW 2751, Australia

guillaume.gibert@inserm.fr, y.leung@uws.edu.au, kj.stevens@uws.edu.au

Address for correspondence:
Guillaume Gibert
INSERM U846
Stem-Cell and Brain Research Institute
18 avenue Doyen Lépine
69675 Bron Cedex
FRANCE
Email: guillaume.gibert@inserm.fr
Phone: +33 (0)4 72 91 34 41
Fax: +33 (0)4 72 91 34 61
URL: http://www.sbri.fr/members/guillaume-gibert.html, http://marcs.uws.edu.au

Abstract

Several puppetry techniques have been recently proposed to transfer emotional facial expressions to an avatar from a user's video. Whereas generation of facial expressions may not be sensitive to small tracking errors, generation of speech-related facial movements would be severely impaired. Since incongruent facial movements can drastically influence speech perception, we proposed a more effective method to transfer speech-related facial movements from a user to an avatar. After a facial tracking phase, speech articulatory parameters (controlling the jaw and the lips) were determined from the set of landmark positions. Two additional processes calculated the articulatory parameters which controlled the eyelids and the tongue from the 2D Discrete Cosine Transform coefficients of the eyes and inner mouth images.

A speech in noise perception experiment was conducted on 25 participants to evaluate the system. Increase in intelligibility was shown for the avatar and human auditory-visual conditions compared to the avatar and human auditory-only conditions, respectively. Depending on the vocalic context, the results of the avatar auditory-visual presentation were different: all the consonants were better perceived in /a/ vocalic context compared to /i/ and /u/ because of the lack of depth information retrieved from video. This method could be used to accurately animate avatars for hearing impaired people using information technologies and telecommunication.

Keywords

Talking head; Auditory-visual speech; Puppetry; Facial animation; Face tracking.

1 Introduction

When interacting with humans, avatars use verbal and non-verbal channels of communication. In order to look realistic, the avatars should be able to replicate human movements. Many researchers have investigated real-time control of an avatar's facial movements from various inputs. Talking heads can replicate accurately visual speech movements from text input. For example, the talking head Baldi produces accurate visible English speech [1] by moving his lips, jaw, velum and tongue using a coarticulation scheme [2]. Early work from Brand [3] developed a method to drive facial animation from audio input. Several kinds of motion and physiological data have also been used to animate avatars. For example, Morishima [4] used electromyographic (EMG) signals from a human face to control the facial expressions of an avatar consisting of a biomechanical model. Other equipment such as Optotrak [5] and Light Scanner [6] had also been used to pilot the facial movements of an avatar in real-time. However, such specialist equipment is expensive and hard to use outside of the lab.

An alternative is video-based tracking. Indeed, many studies have investigated the classification of facial expressions [7] from video. Caridakis and colleagues [8] used this kind of video-based system to extract the user's face and hand movements from video before converting the movement information into high-level representations (categories). Pre-determined movements corresponding to each category were then used to control an avatar mimicking the user's original behaviour. Such an approach does not replicate exactly the person's movements, but mimics them after interpretation. Another way to perform real-time avatar puppetry has also been proposed using the video channel to control the head orientation and the audio channel to control the lips/jaw movements after an Automatic Speech Recognition phase [9]. This method relies on the ASR accuracy to provide speech-related facial movements. Again, the facial movements are pre-determined and do not mimic per se the puppeteer's facial movements.

Direct puppetry methods have been proposed mainly focusing on facial expressions. For instance, face transfer with multi-linear models was proposed by Vlasic and colleagues [10]. Their focus was directed towards visemes and facial expressions while no speech dynamics were taken into account. Direct transfer of facial expressions has also been proposed using correspondence functions between landmarks extracted from tracking and MPEG-4 Facial Animation Parameters (FAP) driving the 3D avatar's facial expressions [11, 12]. Unfortunately, most FAPs are low-level and do not take into account speech-specific gestures [13]. More recently, confederates' head movements and facial expressions were manipulated in real-time during videoconference conversations by tracking them (using Active Appearance Models - AAM) and reconstructing an avatar face [14-16]. Even though the authors reported that the participants did not notice the manipulation, no evaluation of the accuracy of the generated movements was performed. This method has several limitations: first, because the animation parameters were obtained from purely statistical methods, the authors could not manipulate specific behaviours (for instance jaw or lip motion) independently or drive generic avatars using FAPs or articulatory parameters. Second, it requires building an AAM model for the user and the avatar. Third, the avatar's face is cropped around the face and represented in 2.5D (3D reconstruction from 2D data). To overcome some of these issues, Saragih and colleagues [17] proposed a real-time puppetry method using only a single image of the avatar and user. A combined generic-semantic model was used to transfer the puppeteer's facial expressions to any avatar's face. The oral cavity was transferred by copying the user's oral cavity image onto the avatar. The tongue and teeth appearance looked realistic but lighting differences can still be observed between the avatar's texture and the image of the inner mouth area.

As facial expression generation may not be sensitive to small tracking errors, generation of speech-related facial movements could be severely impaired leading to auditory-visual integration issues. Indeed, speech is in essence a multimodal phenomenon. Visual and acoustic modalities are integrated automatically and at a very early stage (neural evidence shows an integration in the first 200 ms after

100 acoustic onset) [18]. A good example of integration is the McGurk effect [19] which is an automatic
101 perceptual phenomenon appearing under incoherent multimodal information (e.g., when confronted
102 with incongruent auditory and visual speech, subjects report hearing a percept different from the
103 acoustic and/or visual signal). An inaccurate transfer of facial motion can modify the perceived
104 sounds; this effect is enhanced in adverse conditions (background noise for example [20]).
105 The present paper describes a new method to mimic directly the user's speech facial movements from
106 a video or a webcam. First, an Active Shape Model (ASM) [21] was built using a corpus of nonsense
107 words (Vowel Consonant Vowel - VCV) that have been manually landmarked. This model was
108 composed of 68 landmarks on the face (jaw, lips, eyes, eyebrows). From this corpus, an articulatory
109 model was also learnt using a guided PCA procedure. This procedure provided a set of semantically
110 different parameters, i.e. each component drives a specific speech articulation. Then, given a video
111 file or webcam video stream, images were captured at 25 Hz. The ASM delivered the position of the
112 landmarks on the user's face for each image. Finally, articulatory parameters were estimated from the
113 landmark positions and sent to the animation module of an avatar. In addition, a separate process
114 generated the animation of the eyelids using linear correspondences between the Discrete Cosine
115 Transform (DCT) coefficients calculated on the eye images and the eyelid articulatory parameters.
116 The same procedure was applied on the inner mouth images and the tongue articulatory parameters to
117 animate the tongue. In fact, the tongue is an important speech articulator. Although most of the time
118 the tongue is occluded, its movements provide useful information for speech perception as shown in
119 [22] where a point-light display that included additional dots on the tongue and the teeth elicited
120 better performance than a display with 'lips only' dots. A speech in noise experiment was conducted
121 on 25 participants to evaluate the quality of speech-related facial movement transfer. It was
122 hypothesized that the facial movement transfer would improve speech perception in adverse
123 conditions. The first section describes the training phase necessary to build the ASM, the articulatory
124 model and the conversion matrices to animate the tongue and the eyelids. The second section

describes the video puppetry system and the different processes involved. Finally, in the third section we describe the perception experiment conducted to evaluate the system.

2 Training Phase

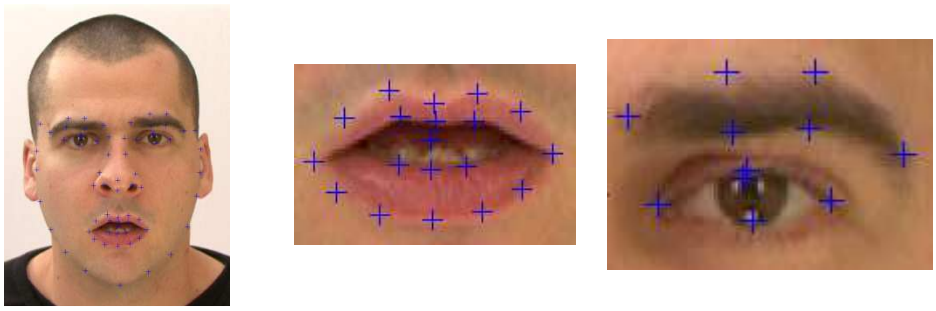
2.1 Material

An Australian English speaker uttered 3 times a series of nonsense words with the following structure Vowel-Consonant-Vowel (VCV). The initial and final vowels of these nonsense words were identical and chosen between /a/, /i/ and /u/ (extreme lip movements). The consonants of these nonsense words were the Australian English consonants /p/, /b/, /m/, /f/, /v/, /t/, /d/, /l/, /w/, /k/, /g/, /s/, /z/, /n/, /j/, /θ/, /ʃ/, /ʒ/, /tʃ/, /dʒ/, /ɹ/. The database was composed of 21 consonants x 3 vowels = 63 nonsense words. This corpus was chosen for two reasons: first, it provided a corpus to build an articulatory model, second it could be used to create stimuli for speech perception experiments. A video (resolution: 720x576 pixels, encoding: 24 bpp, frame rate: 25 fps, codec: dv, interlaced) consisting of a front view of the speaker against a white background was recorded with a SONY HVR-V1P video camera under good lighting conditions. Images were extracted at 25 Hz with the software mencoder (http://www.mplayerhq.hu/). The sound was recorded with an AKG C417 III PP (stereo, 48 kHz, 16 bits) lapel microphone. The sound was also extracted using mencoder software.

2.2 Landmark positions dataset

One complete set of images (every consonant in all symmetrical vocalic contexts) was manually segmented, i.e., the positions of 68 landmarks were selected by hand for each image. This set of landmarks covered the speaker's face and more particularly his speech articulators, i.e., jaw and lip contours (see Figure 1). In fact, twelve landmarks were positioned on the outer lip contour, six on the inner lip contours and nine for the jaw line. Additional landmarks were positioned on the eyebrow contours (six for each eyebrow), the eye contours (five for each eye), the nose contours and the face

contours. Even though increasing the number of landmarks increases the search time, it also improves the mean fit [21]. The database consisted of 3751 segmented images.



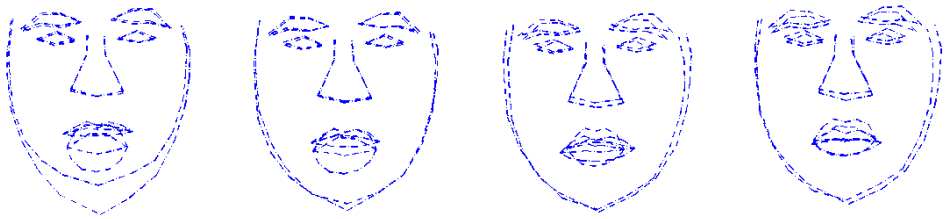
(a) Full face (b) Mouth (c) Eye and eyebrow

Figure 1: Position of the 68 landmarks on the speaker's face for three images extracted from the nonsense word /ipi/. Twelve landmarks were positioned on the outer lips, 6 for the inner lips, 9 for the jaw, 5 for each eye and 6 for each eyebrow.

The annotated database was then cleaned. Whereas selection of the landmark positions on the lips corners is not challenging, consistent selection (over frames) of the landmark positions on the jaw line is not straightforward, and some jitter on the landmark positions may be inserted. In order to remove the movement of the landmarks that were not related to speech motion, an articulatory model was built using the method proposed by [23, 24]. The rigid head motion (translation in x-, y- and z-axes and rotation around the z-axis) was estimated using the landmarks positioned on the eye corners and on the nose tip. Then, the contribution of the speech articulators (lips and jaw in this study) and the eyebrows were iteratively subtracted. This subtraction consisted of an iterative application of Principal Component Analysis (PCA) on subsets of landmarks. The procedure extracted 5 articulatory parameters (see Figure 2):

1. Jaw opening (*jaw1*) using PCA on the jaw position values (35.2% of the global variance);
2. Lip rounding (*lips1*) using PCA on the residual lip position values (16.5% of the global variance);

3. Lip closing (*lips2*) using PCA on the residual lower lip position values (8.2% of the global variance);
4. Lip raising (*lips3*) using PCA on the residual upper lip position values (8.6% of the global variance);
5. Eyebrows raising (*eyebrow*) using PCA on the residual eyebrows position values (4.8% of the global variance).



(a) Jaw1 (b) Lips1 (c) Lips2 (d) Lips3

Figure 2: Variation of the first four articulatory parameters (jaw, lips1, lips2 and lips3) between -3 and +3.

More than 73% of the global variance (after rigid motion extraction) was explained by these 5 articulatory parameters. With the same method, Bailly and colleagues were able to explain more than 95% of the variance of facial movements with six parameters for different subjects and different languages (French, English, German and Arabic) [13]. The parameters they used were the ones described in this paper plus a component for jaw advance and a component for vertical movements of the throat. These two additional parameters explained less than 2% of the variance. The residual movement (27% of the variance) was set to null due to the jitter inserted during the manual segmentation and rotations around x and y axes. In fact, a PCA was applied on the residual movement to ensure it contained 'noise' only. The first component (35.8% of the remaining 27%) corresponded to a jitter of landmark positions along the jaw line. The second component (13.7% of the remaining 27%) corresponded to a jitter of landmark positions along the jaw line and a slight rotation around the y-axis. The third component (10.7% of the remaining 27%) corresponded to a slight rotation around

the y- and x-axes and a jitter of landmark positions along the jaw line. The fourth component (5.4% of the remaining 27%) corresponded to a jitter of landmark positions introduced by the manual segmentation along the jaw line and lip contours. The landmark positions were then reconstructed with the contribution of the speech articulators only. This 'clean' database will be referred to hereafter as the 'gold standard' database.

2.3 Active Shape Models

An Active Shape Model (ASM) [25] consists of 2 sub-models: the *shape model* and the *profile model*. The *profile model* describes the characteristics of the image around each landmark. These characteristics are learnt during the training phase by sampling the area around each landmark for all the images of the training database. For one-dimensional (1D) profiles, the normalized gradient of the gray image intensity of a vector orthogonal to the shape edge at each landmark position is computed. For two-dimensional (2D) profiles, a square region around each landmark is used to compute the gradient of gray image intensity. Whereas the original ASM used 1D profile models, 2D profile models have been shown to improve landmark position accuracy [21]. The *shape model* describes the possible relative position of the landmarks with respect to each other. During the training phase, a PCA is applied to the shapes contained in the training database that defines the average facial shape and the permissible distortions around the average. During the search phase, the *profile model* tries to locate each landmark independently by moving the landmark to the position that best matches the model, and then the *shape model* corrects the suggested locations by constraining their relative positions. These two steps are performed successively until convergence. In this study, ASMs were built using the toolbox STASM developed by Milborrow and Nicolls [21].

The ASM models (1D and 2D profiles) provided with the STASM toolbox were built using a set of neutral faces and mild facial expression images. These models cannot capture accurately speech-related facial movements like jaw opening or lip protrusion. The 'gold standard' database was used to build two ASMs: one with a 1D *profile model* and one with a 2D *profile model*. To evaluate the

ability of the ASMs to model ‘speech’ images, a search was performed on the ‘gold standard’ database. This was an ideal search case because the same database was used for training and test. The reconstruction error for the two ASMs was inferior at half a pixel on average. The 2D profile ASM model performed slightly better than the 1D profile ASM model with an average of 0.0724 and 0.0745 pixels, respectively, and median of 0.0453 and 0.0477 pixels, respectively. These results suggested that the ASMs generated with this toolbox can accurately track ‘speech’ images, i.e., images where lip image profiles can be very different between frames for instance open/closed mouth. The 2D profile ASM model will be used in the following sections.

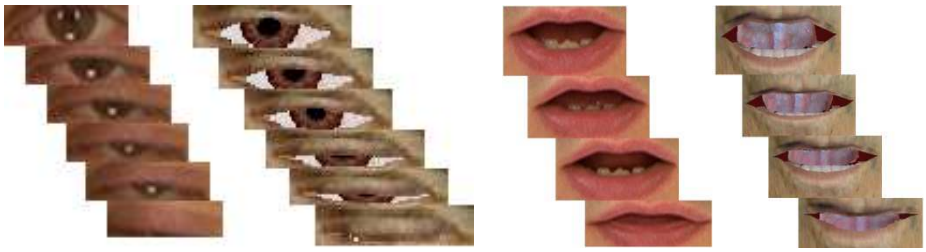
2.4 Eyelid animation training

Cropped images (with a minimal height of 15 pixels) around each eye were created from the landmark positions. The DCT coefficients were extracted for each image converted to greyscale. Only 224 components were kept: 15x15 components in the top left corner of the DCT matrix except the first which is related to the mean value. A quantization (performed on the remaining DCT coefficients) was then performed to select only a small amount of very different images. The percentage of selected frames was 1.7%, i.e., 64 frames for each eye. A manual transcription between these selected images and the articulatory parameter driving the eyelids was then performed (see Figure 3(a) for a set of images). The least squares solution was then determined for the linear correspondence between the DCT coefficients and the articulatory parameter values. This conversion matrix will be used in the online phase to generate the avatar’s blinking pattern.

2.5 Tongue animation training

A similar method was used for the tongue articulator. Cropped images (with a minimal height of 13 pixels) around the inner mouth area were created from the landmark positions. The DCT coefficients were extracted for the red component of each image. The red component was chosen because it conveys the majority of the tongue information without favouring the teeth information. Only 168

components were kept: 13x13 components in the top left corner of the DCT matrix except the first which is related to the mean value. A quantization (performed on the remaining DCT coefficients) was then performed to select only a small number of very different images. The percentage of selected frames was 2.8%, i.e. 106 frames. A manual transcription was conducted between these selected images and the four (out of five) articulatory parameters driving the tongue (see Figure 3(b) for an example). The first articulatory parameter driving the tongue corresponds to the jaw opening and is determined from the landmark positions of the jaw line. The least squares solution was then determined for the linear correspondence between the DCT coefficients and the articulatory parameters values. This conversion matrix will be used in the online phase to generate the avatar's tongue movements.



(a) Eyelid correspondence (b) Tongue correspondence

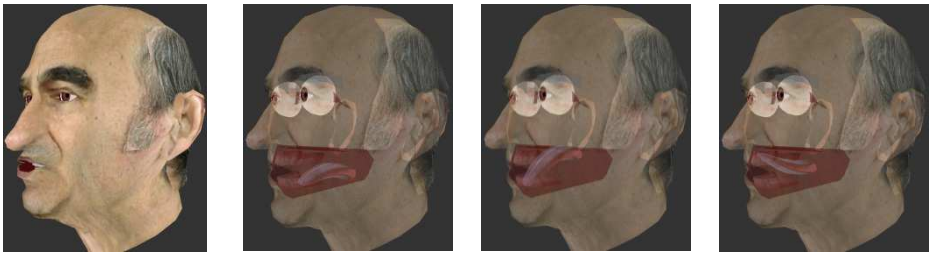
Figure 3: Examples of images of the quantized databases and the corresponding manual implementation with the avatar for the eyelids (a) and the tongue (b).

3 Video puppetry

3.1 Avatar control

The avatar used was a 3D representation of the Australian performance artist Stelarc. This 3D model was driven by articulatory parameters: one for the jaw rotation, 3 for the lips, one for each eyelid and several for facial expressions (surprise, sadness…). Six additional parameters for rotations and translations were also available to control the rigid head motion. The eyes were separate 3D objects which were controlled separately and constituted a visual system. The tongue model was controlled by 5 articulatory parameters (as described in [26]): jaw height (JH), tongue body (TB), tongue dorsum

(TD), tongue tip vertical (TTV) and tongue tip horizontal (TTH). Examples of the maximum variation of some articulatory parameters are shown in Figure 4. 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 during the training phase.



(a) Lips1 (b) TB (c) TD (d) TTV

Figure 4: Examples of the maximum variation of some articulatory parameters driving the avatar. (a) Lips1 corresponds to lip protusion; (b) Tongue body (TB) corresponds to the front-back movement; (c) Tongue dorsum (TD) corresponds to the flattening-bunching movement; (d) Tongue tip vertical (TTV) corresponds to the tongue tip vertical movement.

3.2 Video puppetry animation

The video puppetry animation consisted of several steps: an image was captured from the video stream and then cropped around the face using a face detector (OpenCV implementation of the Viola-Jones algorithm [27]), then the ASM searched the best landmark positions for this image and the jaw and lip articulatory parameters were determined; the eyelids and tongue articulatory parameters were then estimated from the DCT coefficients of cropped images around the eye and oral cavity areas. A diagram of the complete procedure can be found in Figure 5.

3.3 Facial animation

Given a video (different from the ones used during the training phase), images were extracted at 25 Hz. The landmark positions were then determined for each image using the 2D profile ASM learnt on the gold standard database. Articulatory parameter values were then computed from the landmark

positions and the articulatory model (learnt on the gold standard database). This was an optimization step where the best set of parameters was determined to fit the current 2D configuration. Four articulatory parameters were estimated at this stage: jaw1, lips1, lips2, and lips3. These values were then passed to the avatar's shape model which generated a new 3D facial configuration.

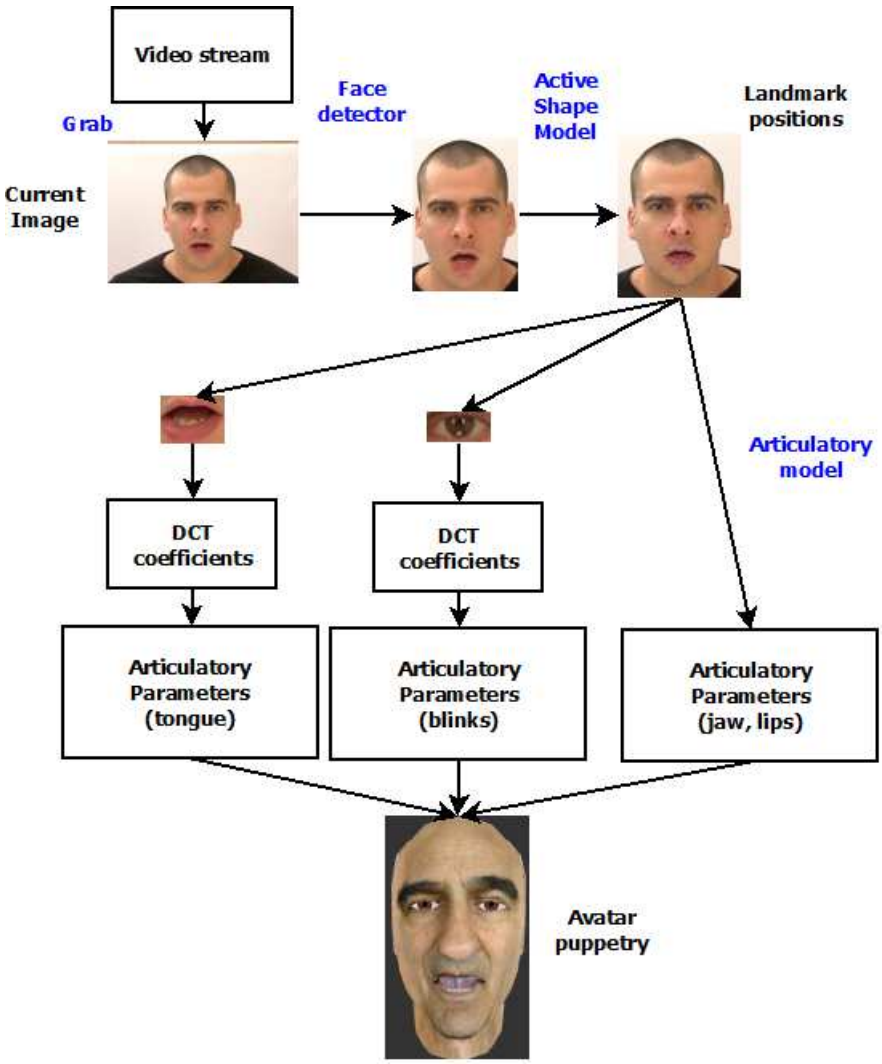


Figure 5: Flowchart of the video puppetry system. An image is captured from the video file, then the ASM delivers the position of the landmarks for the given image. The values of the articulatory parameters (jaw1, lips1, lips2 and lips3) are determined given the articulatory model and the landmark positions. Then, two separate processes determine the values of the articulatory parameters driving the tongue and the eyelids from cropped images around the inner mouth and eye areas. All these parameters are sent to the animation module of the avatar which mimics the user facial configuration.

3.4 Eyelids animation

The current image was cropped around each eye with the help of the landmark positions. The 2D DCT coefficients were then determined and 224 components were kept: 15x15 components in the top left corner of the DCT matrix except the first one. Given the conversion matrix (determined in the training phase) and the values of the DCT coefficients, the articulatory parameters driving the eyelids were then estimated and passed to the avatar's shape model which mimicked the puppeteer's blinking pattern.

3.5 Tongue animation

The tongue was processed in a similar way to the eyelids. The 2D DCT coefficients for the inner mouth area image were calculated. 168 components were kept: 13x13 components in the left top corner of the DCT matrix except the first which is related to the mean value. Given the conversion matrix (determined in the training phase) and the values of the DCT coefficients, the articulatory parameters driving the tongue were estimated and passed to the avatar's shape model which mimicked the puppeteer's tongue movements.

3.6 Results

3.6.1 Eyelids animation

Using the DCT coefficients transfer technique, blink patterns were accurately mimicked onto the avatar. The temporal decomposition of the longest involuntary (~400ms) blink performed by our participant is shown in Figure 6. Each eye was processed separately; asymmetries could be transmitted as displayed in the penultimate image of the series.

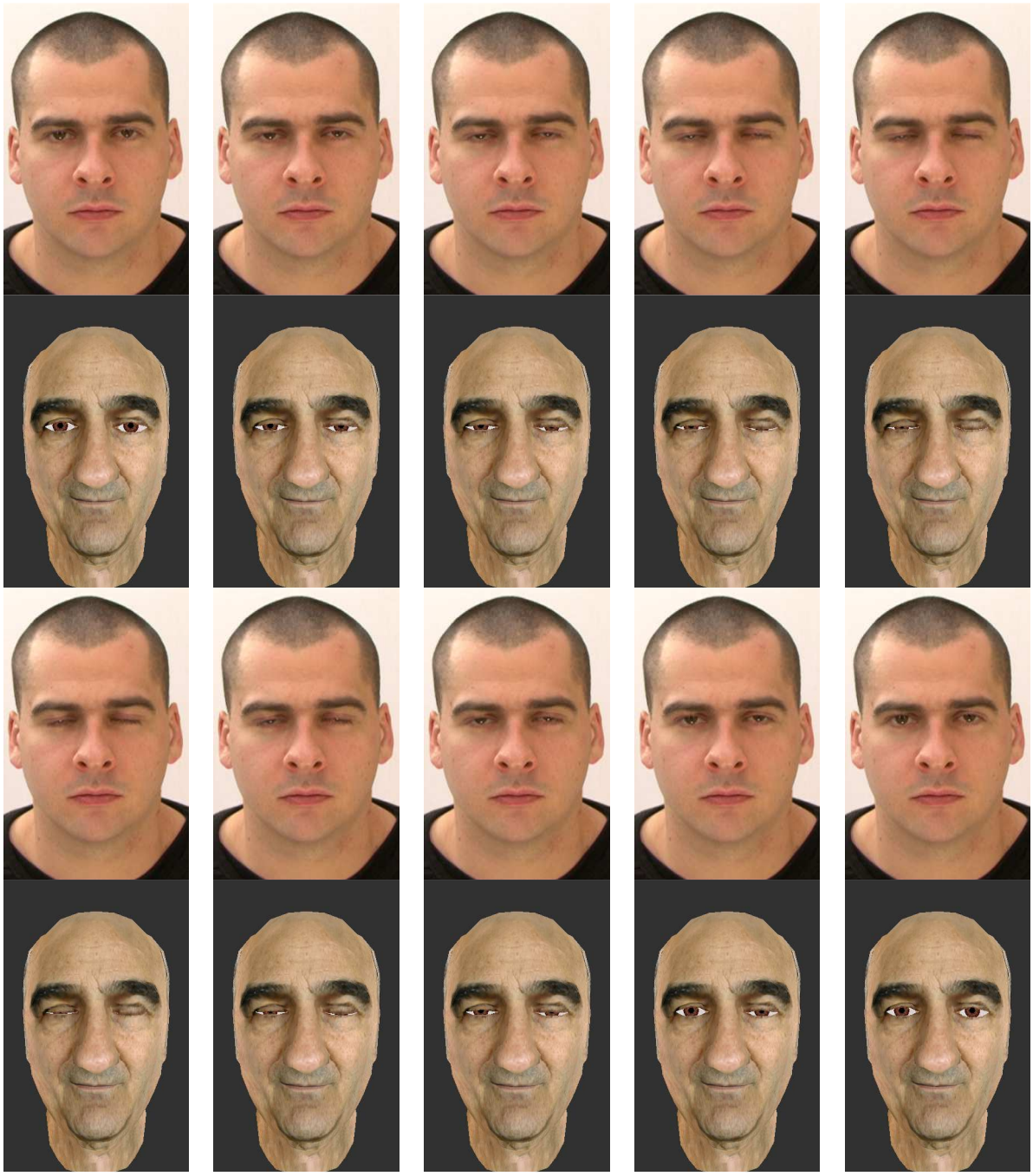


Figure 6: Temporal decomposition (25 Hz) of an involuntary blink. On the even rows, the avatar mimics the puppeteer's blink pattern (odd rows). The asymmetry of the blink pattern is accurately cloned on the avatar's eyelids on the penultimate image.

3.6.2 Tongue animation

The proposed method can transfer accurately the lip and jaw movements from a video. In addition, tongue movements were generated from the 2D DCT coefficients of the oral cavity images. Frames of the utterance of the nonsense word /igi/ are displayed in Figure 7. The middle frame corresponds to the constriction necessary for the production of the consonant /g/. The tongue dorsum movement is accurately recovered from the inner mouth image.

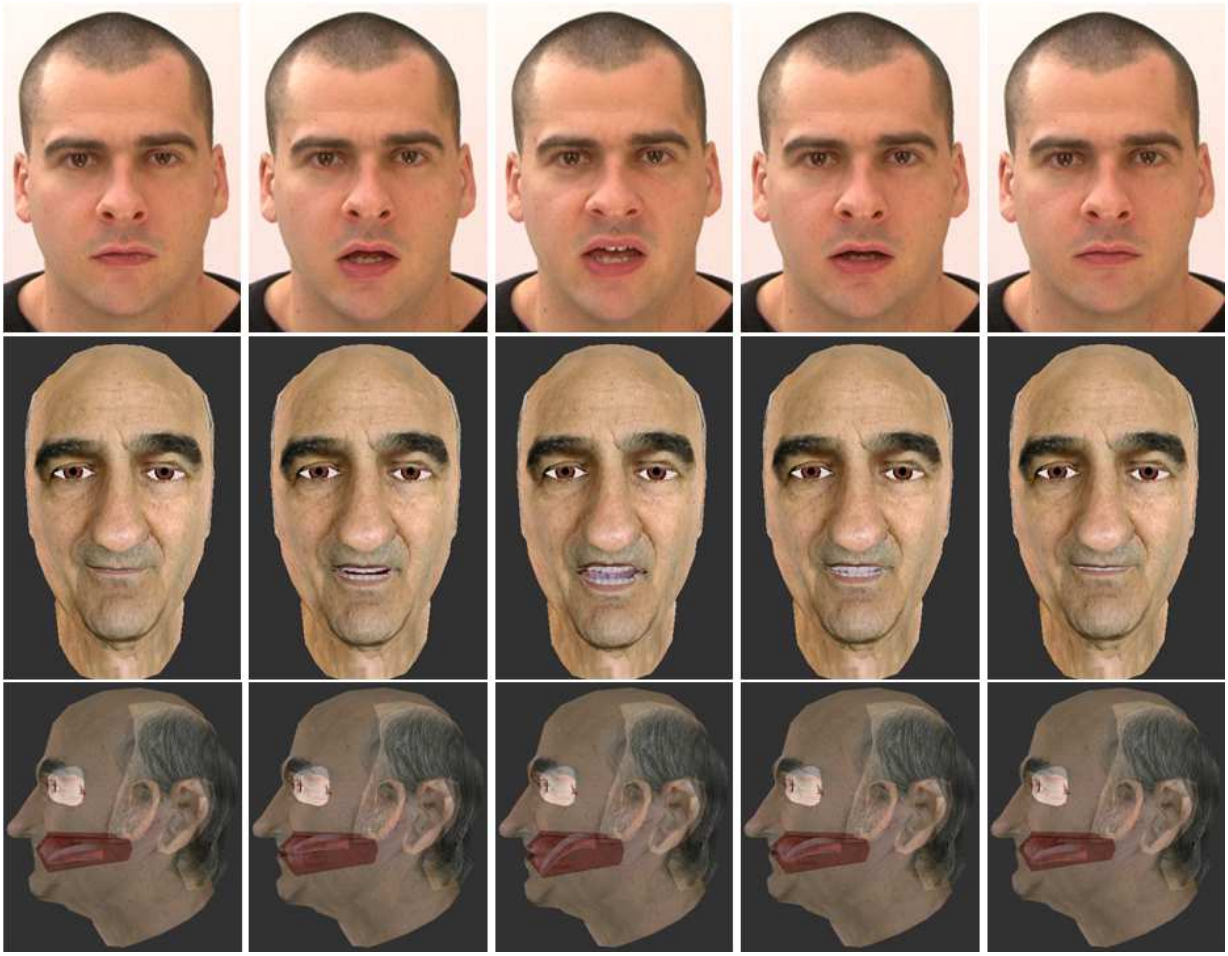


Figure 7: Temporal decomposition of the nonsense word /igi/. On the first row, the speaker utters the nonsense word (front view). In the second and third row, the avatar mimics the speech-related facial movements. The third row represents a side view with a transparent factor of the avatar texture showing the recovered tongue movement from the inner mouth area images.

4 Evaluation

In order to evaluate the movement generation system, a speech in noise experiment was conducted. Visual speech enhances speech perception in noisy conditions [28]. Intelligibility scores are also higher for stimuli displayed using a lip model or face model of a virtual head compared to audio alone [29, 30]. However, if the visual channel is not congruent and/or desynchronized, the percept could be different from the acoustic stimulus [19]. Therefore, this paradigm of evaluation, using a speech perception in noise task, provides an objective measure of the quality of animation transfer.

The independent variables of the experiment were:

- IV1: modality (human audio only – HAO, avatar audio only - AAO, human video - HAV, avatar video - AAV), HAO/HAV-within subjects, AAO/AAV-within subjects. The between subjects factor was Human vs. Avatar. Note that for all the conditions, audio signals from the speaker's video recordings were used, only the visual modality was manipulated;
- IV2: level of noise (Clear, -6dB, -12dB, -18dB) – within subjects, Clear corresponds to the perfect case, i.e. very easy perception task (baseline), -6dB corresponds to a light white noise superimposed on the acoustic signal, i.e. easy perception task, -12dB corresponds to an average level of white noise, i.e. difficult perception task, -18dB corresponds to a loud white noise, i.e. very difficult perception task;
- IV3: vocalic context (/a/, /i/, /u/) – within subjects.

The dependent variable was the identification score, i.e. the number of correct identifications of nonsense words out of the total number of stimuli per condition.

It was hypothesized that HAV and AAV presentations would enhance speech intelligibility compared to HAO and AAO presentation, respectively. This AV enhancement should be smaller for the AAV presentation compared to the HAV because tracking errors and transfer of articulatory parameters through DCT coefficients may impoverish the avatar's speech capabilities. The videos used in the perception experiment were different from the ones used in the training phase. Therefore, tracking

errors may be greater than the ones observed during the training phase. An AV enhancement would reflect efficient facial movement transfer whereas no AV enhancement would reflect poor facial movement transfer. These hypotheses were extended for all levels of noise and it was expected that speech intelligibility differences between auditory-visual and audio only conditions would be greater in conditions of increasing levels of noise. Regarding the vocalic context, it was hypothesized that the identification score would be smaller for the vowel /i/ and /u/ compared to /a/. The main reasons are the lack of depth information (jaw opening movement is easier to track than lip protusion) and the greater amount of information available in the inner mouth area in the /a/ vocalic context compared to /i/ and /u/ giving a better estimation of the tongue movements.

4.1 Methods

4.1.1 Participants

Twenty five first year undergraduate psychology students (19 women, mean participants age 21 years) from the University of Western Sydney participated in this experiment. They were all native Australian English speakers. They received course credit for their participation. All reported normal or corrected-to-normal vision and no hearing loss. This study was approved by the University of Western Sydney Human Research Ethics Committee (H7776).

4.1.2 Stimuli

Videos of VCV nonsense words (not used during the training phase) were segmented on the acoustic stream information. The videos started 400 ms before the acoustic onset of the first vowel and terminated 400 ms after the acoustic offset on the second vowel. The size of each video was 720x576 with a frame rate equal to 25 Hz. The sound track of each video was extracted and four levels of noise were added: Clear, -6dB, -12dB and -18dB. The noise level was computed on the part of the signal containing the nonsense word only, i.e. excluding the first 400 ms and the last 400 ms of signal. The ASM model was applied on the images extracted from each video. Following the procedure described

in Section 3, a set of articulatory parameters for the jaw, lips, tongue, eyebrows and eyelids was generated for each image. Given this set of articulatory parameters an avatar's image was then created. The human and avatar videos were then created from the corresponding set of images and audio tracks. For the HAO and AAO conditions, a static image of the human subject or the avatar (respectively) in resting position (closed mouth) was displayed during the duration of the stimulus.

4.1.3 Procedure

The experiment was conducted in a testing booth. Visual stimuli were displayed on a laptop (Lenovo T500) with a 15'' screen (refresh rate 60 Hz) and audio stimuli were presented through headphones (Senheiser HD650). Participants (seated 0.5 m from a computer screen) were instructed to listen to each stimulus and to identify the nonsense word by clicking on the corresponding labelled button of a graphic user interface using the computer mouse. The labelled buttons consisted of a list of 5 items (e.g., for the /u/ context, the items were /uBu/, /uDu/, /uGu/, /uVu/ and /uZu/). The choice positions were kept constant during each block. All stimuli were presented in a random order by DMDX [31]. Half of the participants (12 subjects) performed the task with the HAV and HAO presentations whereas the other half (13 subjects) perceived the AAV and AAO presentations. An upper limit of time of 4 s on each trial was defined but participants were instructed to respond quickly and to report their first percept. The practice block consisted of 3 stimuli. The experiment comprised of 3 trial blocks of stimuli that were randomized across subjects, one for each vowel /a/, /i/ and /u/. In each block, all levels of noise, modalities and consonants were randomized. Each block was composed of 160 stimuli: 4 (levels of noise) x 2 (modalities) x 5 (consonants) x 2 (items) x 2 (repetitions). Participants could rest in between blocks. The stimulus was played once per trial. After choosing an item, the next stimulus was presented. Viewer responses were recorded using DMDX.

4.2 Results

The identification scores for the different modalities as a function of Signal-to-Noise Ratio (SNR) and collapsed for all vocalic contexts are represented in Figure 8. Consistent with our hypothesis and previous findings, the identification score increased with the SNR in all modality conditions (HAO, AAO, HAV, AAV). The scores were higher for auditory-visual conditions (HAV, AAV) compared to their respective auditory only conditions (HAO, AAO). Two separate 3-way repeated measures ANOVAs were applied for the avatar and the human condition. The repeated factors were modality (HAO or AAO, HAV or AAV), vowel (/a/, /i/, /u/) and level of noise (Clear, -6dB, -12 dB, -18 dB). The results of these ANOVAs are presented in the following section.

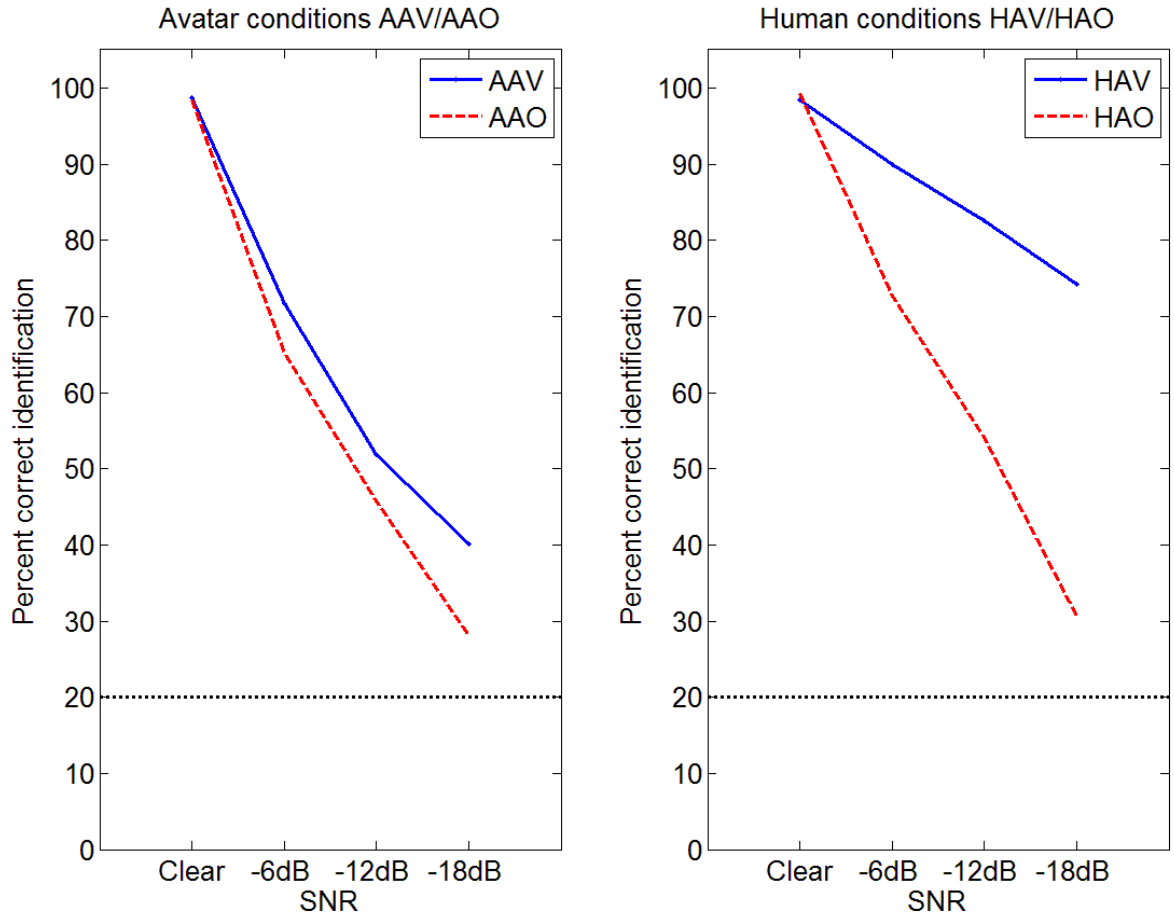


Figure 8: Identification score for all conditions HAO, HAV, AAO, AAV as a function of signal to noise ratio (SNR). The identification score is greater when the SNR is higher for all conditions of presentation. Multimodal presentation (HAV, AAV) elicited higher scores than the corresponding unimodal presentation (HAO, AAO). Note that chance is 20% and represented as a horizontal dashed line.

4.2.1 Avatar condition AAO/AAV

The following results correspond to the avatar condition displayed in auditory only and auditory-visual modalities. There was a significant main effect of modality [$F(1,12)=21.57$, $p=.001$, partial $\eta^2=.64$]. Indeed, the mean intelligibility score was higher for AAV (M = 65.61%, SD = 25.62) than for AAO (M = 59.36%, SD = 30.14). There was a significant main effect of vowel [$F(2,24)=23.34$, $p<.001$, partial $\eta^2=.66$]. Identification score for vowel /a/ (M = 55.12%, SD = 5.81) was significantly greater than for vowel /i/ (M = 45.96%, SD = 7.26) and vowel /u/ (M = 48.88%, SD = 3.69) but no difference was found between vowel /u/ and vowel /i/. There was a significant main effect of level of

noise [$F(3,36)=674.56$, $p<.001$, partial $\eta^2=.98$]. The intelligibility score for Clear (M = 98.59%, SD = 1.87) was greater than for -6dB (M = 68.40%, SD = 6.90) which was greater than for -12dB (M = 48.85%, SD = 7.84) which was greater than for -18dB (M = 34.10%, SD = 9.12). A significant vowel-modality interaction [$F(2,24)=14.57$, $p<.001$, partial $\eta^2=.55$] was found. The difference between vowel /a/ and /u/ was greater in AAV (M = 58.38%, SD = 5.38 and M = 48.38%, SD = 3.52 respectively) than in AAO (M = 51.85%, SD = 4.26 and M = 49.38%, SD = 3.93 respectively) and the difference between vowel /i/ and vowel /u/ was greater in AAO (M = 41.23%, SD = 4.49 and M = 49.38%, SD = 3.93 respectively) than in AAV (M = 50.69%, SD = 6.41 and M = 48.38%, SD = 3.52 respectively).

4.2.2 Human condition HAO/HAV

The following results correspond to the human condition displayed in auditory only and auditory-visual modalities. There was a significant main effect of modality [$F(1,11)=174.44$, $p<.001$, partial $\eta^2=.94$]. The identification score was significantly higher for HAV (M = 86.25%, SD = 10.37) than for HAO (M = 64.20%, SD = 29.02). There was a significant main effect of vowel [$F(2,22)=4.72$, $p=.02$, partial $\eta^2=.30$]. The identification score was significant higher for vowel /a/ (M = 61.46%, SD = 8.20) than for vowel /i/ (M = 58.75%, SD = 13.47). There was a significant main effect of noise [$F(3,33)=518.05$, $p<.001$, partial $\eta^2=.98$]. The identification score was greater for Clear (M = 98.89%, SD = 1.27) than for -6dB (M = 81.25%, SD = 9.86) which was greater than for -12dB (M = 68.33%, SD = 16.00) which was greater than -18dB (M = 52.43%, SD = 23.69). A significant vowel-modality interaction [$F(2,22)=26.46$, $p<.001$, partial $\eta^2=.85$] was found. The difference between vowel /i/ and vowel /u/ was greater in HAO (M = 46.25%, SD = 4.43 and M = 53.58%, SD = 4.46 respectively) than in HAV (M = 71.25%, SD = 4.35 and M = 67.08%, SD = 3.80 respectively).

4.2.3 Confusion matrices

The confusion matrices for responses to HAV and AAV modalities are provided in Table 1.

In /a/ vocalic context, the avatar presentation was weaker than the human one for the consonant /b/ showing lip closure issues for some stimuli. Most of the errors corresponded to the perception of /d/ and /v/ consonants in this case. It is worth noting that /d/ and /g/ were well perceived in HAV and AAV, i.e. transfer of tongue movement was efficient in this case. Another difference between HAV and AAV was for /v/: the identification score was almost perfect for HAV but some errors occurred for AAV with mainly perception of /d/ and /z/ consonants. The transfer of the tongue and lip movements was not optimal in this case.

In /i/ vocalic context, lip closure issues arose for /b/ which tended to be more misperceived in AAV than in HAV with mainly /d/ and /v/ responses. More errors appeared between /d/ and /g/ in this vocalic context than in /a/. Perception of /d/ and /z/ consonants for /v/ presentation was observed as in /a/ vocalic context.

In the /u/ vocalic context, there was even more lip closure issues with more than half of the responses to /b/ not perceived as /b/. In this vocalic context, there is a difficulty in transferring accurately the tongue movements related to the consonant /g/. Indeed, /g/ was perceived as /d/ or /z/.

Table 1: Confusion matrices for responses to HAV and AAV modalities separated by vocalic context and collapsed over all SNR and subjects. For a given consonant stimulus (row), columns correspond to the number of perceived consonants among the 5 possible ones. Maximum score is 192 for HAV corresponding to 2 repetitions x 12 subjects x 4 levels of noise x 2 items and 208 for AAV corresponding to 2 repetitions x 13 subjects x 4 levels of noise x 2 items.

		/a/					**/i/**					**/u/**				
		b	**d**	**g**	**v**	**z**	**b**	**d**	**g**	**v**	**z**	**b**	**d**	**g**	**v**	**z**
HAV	**b**	**187**	4	1	0	0	**185**	4	1	2	0	**156**	6	2	28	0
	d	1	**141**	48	0	2	0	**145**	23	1	23	1	**125**	59	0	6
	g	0	3	**169**	0	20	0	19	**172**	1	0	1	34	**147**	0	9
	v	1	0	0	**187**	4	1	1	1	**189**	0	0	0	1	**189**	2
	z	0	11	4	37	**140**	0	1	1	26	**164**	0	2	1	1	**188**

	b	**149**	21	3	28	7	**143**	26	7	25	7	**84**	27	23	52	22
	d	3	**148**	36	6	15	3	**133**	49	7	16	4	**154**	19	4	27
AAV	**g**	1	12	**170**	1	24	4	39	**152**	2	11	1	56	**79**	12	60
	v	12	22	10	**137**	26	13	38	21	**110**	26	11	14	14	**130**	39
	z	1	11	24	17	**155**	1	12	43	31	**121**	1	7	9	9	**182**

4.2.4 Discussion

The auditory-visual presentation is significantly more intelligible than the auditory-only presentation for both the human and the avatar conditions. This result shows that our system is able to transfer speech-related facial movements accurately enough to provide a benefit for speech perception in adverse conditions. The increase in intelligibility is smaller for the avatar compared to the human condition as hypothesized. We can use the relative visual contribution metric proposed by Ouni and colleagues [32] to estimate the quality of the animated talker compared to the natural face. The relative visual contribution metric was defined as follows:

$$C_{RV} = 1 - (C_N - C_S)/(1 - C_A)$$

where C_N corresponds to the HAV intelligibility scores, C_S corresponds to the AAV intelligibility scores and C_A corresponds to the average of AAO and HAO intelligibility scores. The relative visual contribution was equal to 0.48 for -6dB condition, 0.44 for -12dB and 0.53 for -18dB condition. Animated with the proposed transfer method, the avatar reached around 50% of the visual performance of the human. For comparison, talking heads (driven by richer information, i.e. list of phonemes and their durations) can reach more than 80% of the visual performance of a natural face [32]. Jitter in the tracking step, lack of depth information and approximation in the DCT to the articulatory parameter conversion step could explain this result. In order to improve the jitter inserted while tracking, more sophisticated tracking algorithms could be used and the DCT-to-articulatory-parameters conversion could be enhanced by using more varied data during the training phase.

As hypothesized, perception of consonants in /a/ vocalic context was easier than in /i/ and /u/ vocalic contexts. The lack of depth information with the video input may explain this result. In fact, the articulatory model only partially recovered the lip rounding gesture. Using an additional depth sensor (such as the Kinect™ sensor) or fitting a generic 3D model using ASM or AAM should improve the articulatory model and the recovering of the lip rounding gesture.

5 Conclusions and Perspectives

A new method to control an avatar's facial gestures from the video stream of a person speaking (the puppeteer) has been described. Rather than focusing on facial expressions, the proposed technique focused on visual speech articulation and extracted articulatory parameters from landmark positions estimated from each grabbed image. Contrary to most approaches already described in the literature, the eyelids and the tongue were animated. This was performed by separately using the 2D DCT coefficients of images cropped around the eye and the oral cavity areas, respectively. The animation of all the visible (and partially occluded) speech articulators (jaw, lips, and tongue) was efficiently transferred from a human to an avatar. To evaluate the accuracy of the current method, a speech in noise identification experiment was conducted. This evaluation method provided an objective measure of the quality of movement transfer compared to qualitative self-report questionnaires used to assess the transfer of emotional facial expressions proposed in the literature. This experiment consisted of nonsense word audiovisual stimuli with several levels of noise in the acoustic channel (clear, 0dB, -6dB and -12dB) presented to participants. The participants' task was the identification of the perceived nonsense words. The identification scores were compared with the puppeteer's video providing a baseline. This method of evaluation provided an ecologically- valid framework as avatar videos were compared to the corresponding human videos. This method could be used by other facial movements transfer techniques to assess quantitatively the quality of transfer.

To improve the accuracy of the method, the next step will be to use the affordable Kinect™ depth sensor. In addition to the colour image, a depth map can be acquired providing more precise 3D

information than fitting a 3D generic mesh model using ASM or AAM. It would enhance the tracking algorithm and the estimation of the articulatory parameters. It would be interesting to evaluate with the same method two different approaches to transfer tongue movements: the method described in this paper and the transfer by copying/warping the inner-mouth area images of the puppeteer on the avatar.

Using just a webcam, the present method may be used to enhance interactions in virtual worlds by providing accurate facial movements to the interlocutor. Generally, ASM are relatively slow to converge. Recently, a hierarchical ASM working in real-time was proposed [33]. Another way to improve the convergence would be to parallelize the landmark localization which is performed sequentially and represents the bottleneck of the process. The system could be used with generic tracking models ASM or AAM (learnt on a large image dataset such as the CMU Multi-PIE Face database [34]) to determine the landmark positions. A generic articulatory model could also be used but a personal one would provide more accurate movement transfer. This system would increase the interaction realism in 3D environments. Interesting assistive technologies for hearing impaired people could be developed around this technique; for instance, 3D movies could be 'fully dubbed' by manipulating not only the voice but also the facial movements of the 3D characters giving access to lipreading. On the research side, 'super' Wizard of Oz setups could be built using a confederate's speech-related facial movements. Manipulation (for instance adding delay or degrading) of chosen articulatory parameters could be used to investigate, for example, the conditions that give rise to the uncanny valley phenomenon.

6 Acknowledgements

We thank James Heathers for manually segmenting the images. This work was supported by the Thinking Head project, a *Special Initiative* scheme of the Australian Research Council and the National Health and Medical Research Council (TS0669874) [35] and by the SWoOZ project (ANR 11 PDOC 019 01).

7 Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.specom.2012.07.001.

8 References

[1] S. Ouni, D. W. Massaro, M. M. Cohen, K. Young, and A. Jesse, "Internationalization of a Talking Head," presented at the International Congress of Phonetic Sciences (ICPhS'03), Barcelona, Spain, 2003.

[2] D. W. Massaro, *Perceiving Talking Faces, From Speech Perception to a Behavioral Principle*: MIT Press, 1998.

[3] M. Brand, "Voice puppetry," presented at the Proceedings of the 26th annual conference on Computer graphics and interactive techniques, 1999.

[4] S. Morishima, "Face analysis and synthesis - For duplication expression and impression," *Ieee Signal Processing Magazine*, vol. 18, pp. 26-34, May 2001.

[5] T. Kuratate, H. Yehia, and E. Vatikiotis-Bateson, "Kinematics-based synthesis of realistic talking faces," in *International Conference on Auditory-Visual Speech Processing*, Terrigal, Sydney, Australia, 1998, pp. 185-190.

[6] T. Weise, H. Li, L. Van Gool, and M. Pauly, "Face/Off: Live Facial Puppetry," in *Eighth ACM SIGGRAPH / Eurographics Symposium on Computer Animation* New Orleans, LA, USA, 2009.

[7] C. Chibelushi and F. Bourel, "Expression Recognition: A Brief Tutorial Overview," in *CVonline: On-Line Compendium of Computer Vision*, 2003.

[8] G. Caridakis, A. Raouzaiou, E. Bevacqua, M. Mancini, K. Karpouzis, L. Malatesta, and C. Pelachaud, "Virtual Agent multimodal mimicry of humans," *Lang. Resources & Evaluation*, vol. 41, 2007.

[9] S. Gallou, G. Breton, R. Seguier, and C. Garcia, "Avatar Puppetry Using Real-Time Audio and Video Analysis," presented at the Proceedings of the 7th international conference on Intelligent Virtual Agents, Paris, France, 2007.

[10] D. Vlasic, M. Brand, H. Pfister, and J. Popovic, "Face transfer with multilinear models," *ACM Trans. Graph.*, vol. 24, pp. 426-433, 2005.

[11] J. Allbeck, N. Badler, T. Bickmore, C. Pelachaud, A. Safonova, R. Baptista Queiroz, A. Braun, J. Moreira, M. Cohen, S. Musse, M. Thielo, and R. Samadani, "Reflecting User Faces in Avatars," in *Intelligent Virtual Agents*. vol. 6356, ed: Springer Berlin / Heidelberg, 2010, pp. 420-426.

[12] G. Fanelli and M. Fratarcangeli, "A Non-Invasive Approach for Driving Virtual Talking Heads from Real Facial Movements," in *3DTV Conference, 2007*, 2007, pp. 1-4.

[13] G. Bailly, M. Berar, F. Elisei, and M. Odisio, "Audiovisual Speech Synthesis " *International Journal of Speech Technology*, vol. 6, pp. 331-346, 2003.

[14] S. M. Boker, J. F. Cohn, B. J. Theobald, I. Matthews, T. R. Brick, and J. R. Spies, "Effects of damping head movement and facial expression in dyadic conversation using real-time facial expression tracking and synthesized avatars," *Philosophical Transactions of the Royal Society B-Biological Sciences*, vol. 364, pp. 3485-3495, Dec 12 2009.

[15] B.-J. Theobald, I. A. Matthews, J. F. Cohn, and S. M. Boker, "Real-time expression cloning using appearance models," presented at the Proceedings of the 9th international conference on Multimodal interfaces, Nagoya, Aichi, Japan, 2007.

[16] B. J. Theobald, I. Matthews, M. Mangini, J. R. Spies, T. R. Brick, J. F. Cohn, and S. M. Boker, "Mapping and Manipulating Facial Expression," *Language and Speech*, vol. 52, pp. 369-386, 2009.

[17] J. M. Saragih, S. Lucey, and J. F. Cohn, "Real-time avatar animation from a single image," presented at the Automatic Face and Gesture Recognition, 2011.

[18] J. Besle, A. Fort, C. Delpuech, and M. H. Giard, "Bimodal speech: early suppressive visual effects in human auditory cortex," *European Journal of Neuroscience*, vol. 20, pp. 2225-2234, Oct 2004.

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

[20] T. R. Jordan and P. C. Sergeant, "Effects of facial image size on visual and audio-visual speech recognition," in *Hearing by eye II*, R. Campbell, B. Dodd, and D. Burnham, Eds., ed: Psychology Press, 1998, pp. 155-176.

[21] S. Milborrow and F. Nicolls, "Locating Facial Features with an Extended Active Shape Model," in *European Conference on Computer Vision*, Marseille, France, 2008, pp. 504-513.

[22] 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.

[23] 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.

[24] 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.

[25] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active Shape Models-Their Training and Application," *Computer Vision and Image Understanding*, vol. 61, pp. 38-59, 1995.

[26] P. Badin and A. Serrurier, "Three-dimensional linear modeling of tongue: Articulatory data and models," presented at the 7th International Seminar on Speech Production, Belo Horizonte, Brazil, 2006.

[27] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2001, pp. I511-I518.

[28] 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.

[29] C. Benoit and B. Le Goff, "Audio-visual speech synthesis from French text: Eight years of models, designs and evaluation at the ICP," *Speech Communication*, vol. 26, pp. 117-129, Oct 1998.

[30] D. Massaro, M. M. Cohen, H. Meyer, T. Stribling, C. Sterling, and S. Vanderhyden, "Integration of Facial and Newly Learned Visual Cues in Speech Perception," *American Journal of Psychology*, vol. 124, pp. 341-354, Fal 2011.

[31] K. I. Forster and J. C. Forster, "DMDX: a windows display program with millisecond accuracy," *Behav Res Methods Instrum Comput*, vol. 35, pp. 116-24, Feb 2003.

[32] S. Ouni, M. M. Cohen, H. Ishak, and D. W. Massaro, "Visual Contribution to Speech Perception: Measuring the Intelligibility of Animated Talking Heads," *Eurasip Journal on Audio Speech and Music Processing*, 2007.

[33] S. W. Lee, J. Kang, J. Shin, and J. Paik, "Hierarchical active shape model with motion prediction for real-time tracking of non-rigid objects," *Iet Computer Vision*, vol. 1, pp. 17-24, Mar 2007.

[34] R. Gross, I. Matthews, J. F. Cohn, T. Kanade, and S. Baker, "Multi-PIE," *Image and Vision Computing*, vol. 28, pp. 807-813, 2010.

[35] 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," ed: ARC/NH&MRC Special Initiatives, TS0669874, 2006-2011.