Analysis and synthesis of the three-dimensional movements of the head, face, and hand of a speaker using cued speech

Guillaume Gibert, Gérard Bailly,[a] Denis Beautemps, and Frédéric Elisei
Institut de la Communication Parlée (ICP), UMR CNRS 5009, INPG/U3, 46, av. Félix Viallet—38031 Grenoble, France

Rémi Brun
Attitude Studio SA, 50, avenue du Président Wilson—93214 St Denis-la-Plaine, France

(Received 19 July 2004; revised 9 May 2005; accepted 9 May 2005)

In this paper we present efforts for characterizing the three dimensional (3-D) movements of the right hand and the face of a French female speaker during the audiovisual production of cued speech. The 3-D trajectories of 50 hand and 63 facial flesh points during the production of 238 utterances were analyzed. These utterances were carefully designed to cover all possible diphones of the French language. Linear and nonlinear statistical models of the articulations and the postures of the hand and the face have been developed using separate and joint corpora. Automatic recognition of hand and face postures at targets was performed to verify *a posteriori* that key hand movements and postures imposed by cued speech had been well realized by the subject. Recognition results were further exploited in order to study the phonetic structure of cued speech, notably the phasing relations between hand gestures and sound production. The hand and face gestural scores are studied in reference with the acoustic segmentation. A first implementation of a concatenative audiovisual text-to-cued speech synthesis system is finally described that employs this unique and extensive data on cued speech in action. © *2005 Acoustical Society of America.* [DOI: 10.1121/1.1944587]

PACS number(s): 43.72.Ja [DOS] Pages: 1–XXXX

I. INTRODUCTION

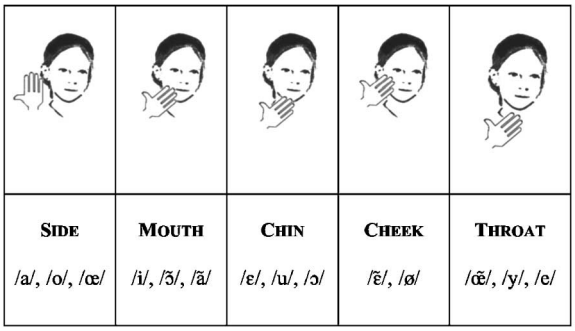
Speech articulation has clear visible consequences. When a person speaks, the movements of the jaw, the lips, and the cheeks are immediately visible. However, the movements of the underlying organs that shape the vocal tract and the sound structure (larynx, velum, and tongue) are not so visible: tongue movements are correlated with visible movements ($R \sim 0.7$) (Kuratate *et al.*, 1999; Yehia *et al.*, 1998; Jiang *et al.*, 2000), but this correlation is insufficient for recovering essential phonetic cues such as place of articulation (Bailly and Badin, 2002; Engwall and Beskow, 2003).

People with hearing impairment typically rely heavily on speech reading based on visual information of the lips and face. However, speech reading alone is not sufficient due to the lack of information on the place of tongue articulation and the mode of articulation (nasality or voicing) as well as to the ambiguity of the lip shapes of some speech units (visemes as [u] versus [y]). Indeed, even the best speech readers do not identify more than 50 percent of phonemes in nonsense syllables (Owens and Blazek, 1985) or in words or sentences (Bernstein *et al.*, 2000). This performance depends on various factors but remains quite far from hearing subjects. The highly trained deaf subjects in the Uchanski *et al.* experiments (1994) obtained mean scores varying from 21% to 62% with lip reading alone, depending on sentence predictability, whereas scores from 78% to 97% are obtained with the help of Cued Speech.

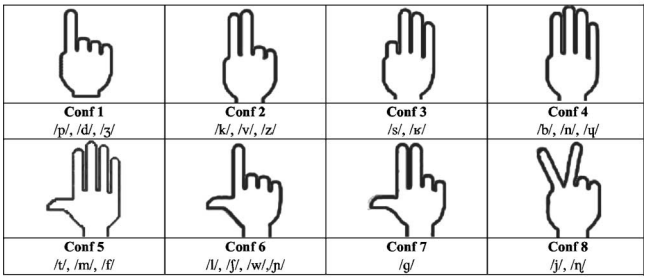
Cued Speech (CS) was designed to complement speech reading. Developed by Cornett *et al.* (1967; 1992) and adapted to more than 50 languages (Cornett, 1988), this system is based on the association of speech articulation with cues formed by the hand. While speaking, the cuer[1] uses one of his/her hand to point out specific positions on the face (indicating a subset of vowels) with a hand shape (indicating a subset of consonants). The French CS (FCS) system is described in Fig. 1. Numerous studies have demonstrated the drastic increase of intelligibility provided by CS compared to speech reading alone (Nicholls and Ling, 1982; Uchanski *et al.*, 1994) and the effective facilitation of language learning using FCS (Leybaert, 2000; Leybaert 2003).

A large amount of work has been devoted to CS perception, but few works have provided insights in the CS production. Attina and colleagues (2002; 2004, 2003a) studied the hand movements of a FCS cuer and their phasing relations with visible (notably lip area) and audible speech. They used a corpus of nonsense words ($[\mathrm{CaCV_1CV_2CV_1}]$ sequences with $\mathrm{C} \in [\mathrm{m,p,t}]$ and V_1 and $V_2 \in [\mathrm{a,i,u},\phi,\mathrm{e}]$) and they observed an average advance of 200 ms for the beginning of the hand gesture with respect to the acoustic realization of the CV syllable, and they also observed the hand target position was reached quasisynchronically with the acoustic consonantal onset of the CV syllable. This confirmed the *ad hoc* rules retained by Duchnovski *et al.* (2000) for their system of automatic generation

[a]Contact Gérard Bailly, ICP, INPG, 46 Avenue Félix Viallet, 38031 Grenoble Cédex 01, France. Electronic mail: bailly@icp.inpg.fr; Phone: 33 + 476 57 47 11; fax: 33 + 476 57 47 10.



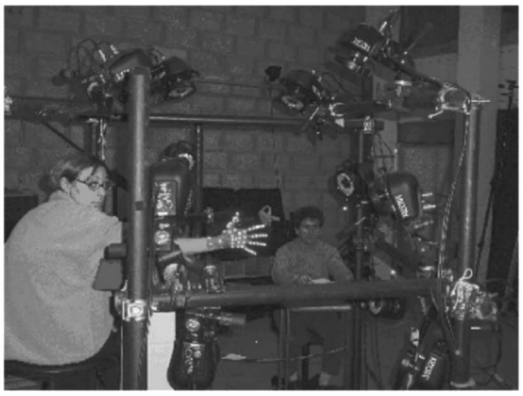
(a) the 5 hand placements. Side is also used for a consonant followed by another consonant or a schwa.



(b) hand shapes. Conf 5 is also used for a vowel not preceded by a consonant.

FIG. 1. French cued speech system.

(a) Position of the retroreflective markers



(b) Experimental setting for capturing the hand motion in free space

FIG. 2. Motion capture experiment.

of CS for English. The authors concluded in a "topsy-turvy vision of Cued Speech" that hypothesizes that the hand placement (i.e., the reached hand target) first gives a set of possibilities for the vowel, the lips then delivering the uniqueness of the solution. Rules for hand movement based on this principle were integrated in a first 2-D audiovisual Cued Speech synthesizer delivering Cued Speech from text (Attina *et al.* 2003b).

In the study we extend this pioneering work to the complete characterization of the 3-D movements of the head, face, and hand. We characterize here the cued speech production of complete utterances. Section II is dedicated to the description of our experimental design for collecting massive motion capture data with high temporal and spatial precision. In Sec. III we describe how motion capture data are regularized using statistical shape models for the face and hand built using selected data. In Sec. IV we further describe the gestural scores we built from motion capture data in order to (a) verify *a posteriori* that the cuer has effectively produced the hand shapes and placements she has to do for complementing speech production (b) study phasing relations of hand placements and hand shapes with the speech signal. In Sec. V we sketch the first version of an audiovisual concatenative text-to-cued speech synthesis system built using the resources of this study.

II. MOTION CAPTURE DATA

We recorded the 3-D positions of 113 retroreflective markers glued on the hands and face of the subject, a skilled cuer who has a daily practice of FCS with relatives, using a Vicon® motion capture system with 12 cameras [see Fig. 2(a)]. The system delivers the 3-D positions of candidate markers at 120 Hz. Recordings and ground truth data processing were performed at Attitude Studio. Further software was provided to assist users in collecting coherent 3-D trajectories and deleting outliers. Note that this tedious semiautomatic task was not error-free. Two different settings of the cameras enabled us to record three corpora.

(i) Corpus 1—*hand only*: the cuer produced all possible transitions between eight hand shapes in free space, with each hand shape corresponding to a subset of consonants [see Fig. 2(b)].
(ii) Corpus 2—*face and audio*: the cuer uttered without cueing, visemes of all isolated French vowels and all consonants in symmetrical context VCV, where V is one of the extreme vowels [a], [i], or [u]. This corpus in similar to the one usually used at ICP for facial cloning (Badin *et al.* 2002)
(iii) Corpus 3—*hand, face, and audio*: the cuer uttered 238 sentences, carefully selected to contain all French diphones. This corpus has more than 200 000 frames in total and was used for FCS recognition and synthesis.

All productions were also videotaped using a camera placed approximately 4 m in front of the cuer. The content of the corpus was delivered sentence by sentence to the cuer by a prerecorded acoustic prompt. This content was available to her several weeks in advance. She was instructed to listen to the acoustic prompt and repeat the utterances aloud as if she was cueing them for a deaf partner standing just behind the video recorder.

For corpus 1, cameras were placed all around the hand in order to gather 3-D positions of the markers even at extreme retracted positions of the fingers. Another camera setting—with a larger working space—was used for the second and third corpora. Corpora 1 and 2 were used to build statistical models of the hand and face movements separately. The models were then used to recover missing data in corpus 3, where the face was partially blocked by the hand, and *vice versa*.

III. ARTICULATORY MODELS OF THE FACE AND HAND

Building statistical models from raw motion capture data has several technological and scientific motivations.

The first technological motivation was to provide a way to clean up the data automatically: the semiautomatic labeling of 3-D trajectories is usually very costly and quite time consuming. The second technological motivation was to ease the work of the infographist, who will be responsible for adjusting the movements of a predefined character or avatar to these raw data: inverse kinematics is more effective when dealing with clean target trajectories.

The scientific motivations concern (1) the study of production of FCS and (2) the coordination between acoustics, face, and hand movements during cued speech production.

A. Face

The basic methodology developed at ICP for cloning speech articulation has already been applied to different speech organs such as the face (Revéret *et al.*, 2000) and the tongue (Badin *et al.*, 2002) and to different speakers (Bailly *et al.*, 2003). From raw motion data, we estimated and subtracted iteratively the elementary movements of segments (lips, jaw,…) known to drive facial motion. These elementary movements were estimated using a Principal Component Analysis (PCA) performed on pertinent subsets of flesh points (e.g., points on the jaw line for the estimation of jaw rotation and protrusion, lip points for lip protrusion/retraction gesture).

This basic methodology was previously applied to quasistatic heads. Since the head was free to move in corpora 2 and 3, we need to solve the problem of the repartition of the variance of the positions of the markers placed on the throat between head and face movements. This problem was solved in three steps.

(1) An estimation of the head movement using the hypothesis of a rigid motion of markers placed on the nose and forehead. A principal component analysis of the 6 parameters of the rototranslation extracted for corpus 3 was then performed and the *nmF* (stands for *number of free movements of the head*) first components were retained as control parameters for the head motion.
(2) Facial motion cloning subtracting the inverse rigid motion of the full data. Only *naF* (stands for *number of free articulatory movements of the face*) components were retained as control parameters for the facial articulation.
(3) Throat movements were considered to be equal to head movements weighted by factors (parameters *wmF* below) less than one. A joint optimization of these weights and the directions of the throat deformations was then performed keeping the same values for the *nmF* and *naF* predictors for each frame.

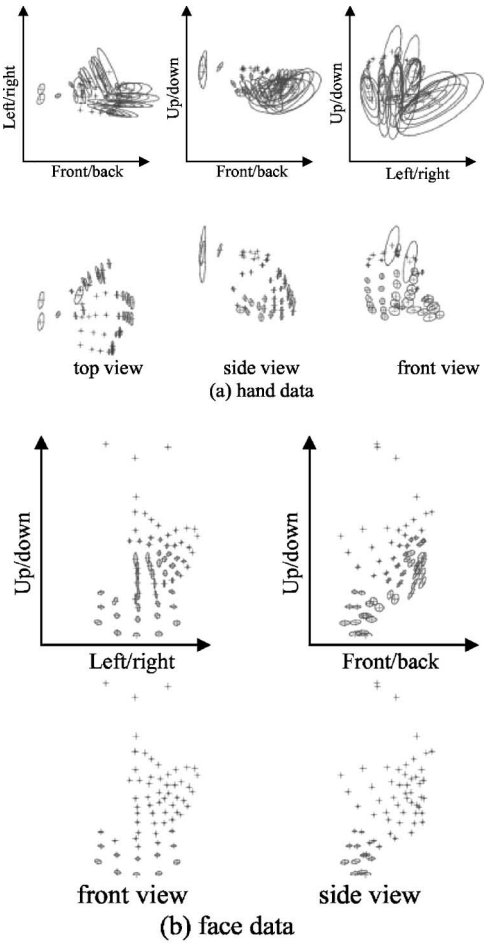


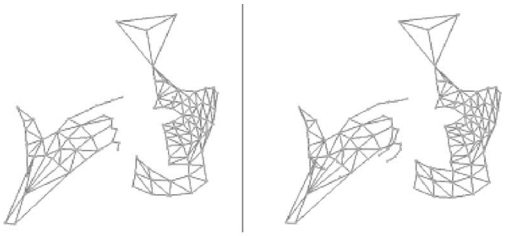
FIG. 3. Dispersion ellipses of the hand and face data (displayed relative to the mean configuration). Top: raw data. Bottom: residual. Both residuals are computed, taking into account both the motion of the segment in space and intrinsic motion.

These operations were performed using facial data from corpus 2 and 3 with all markers visible. A simple vector quantization that guaranteed a minimum 3-D distance between selected training frames (2 mm) was performed before modeling. This pruning step provided statistical models with conditioned data.

The final algorithm for computing the 3-D positions *P3DF* of the 63 face markers of a given frame is:

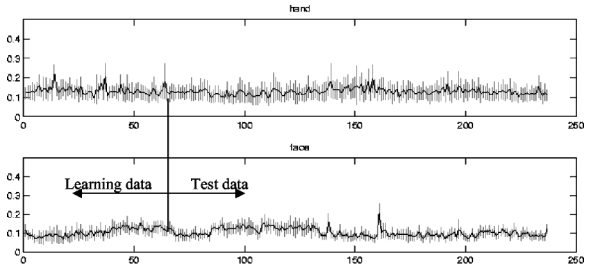
```
mvt = mean_mF+pmF*eigv_mF;

P3D = reshape(mean_F+paF*eigv_F,3,63);

for i=1:63

 M = mvt.*wmF(:,i);

 P3DF(:,i) = Rigid_Motion(P3D(:,i),M);
end
```

where *mvt* is the head movement controlled by the *nmF*



(a) incomplete motion capture data *(b) reconstruction*

FIG. 4. Reconstruction of a FCS frame. Part of the throat and fingers have not been captured by the motion tracking system but have been reconstructed properly by the face and hand models.



(a) learning the models with only the first 68 sentences

FIG. 5. Mean and standard deviation of the mean reconstruction error for each sentence processed by the hand (top) and face (bottom) models. The models were built using the 68 first sentences as learning data. The peak errors for reconstructions are mainly due to false labeling of raw motion data.

parameters *pmF*, *M* is the movement weighted for each marker (equal to 1 for all face markers, less than 1 for markers on the throat) and *P3D* are the 3-D positions of the markers without head movements controlled by *naF* parameters *paF*.

B. Hand

Building a statistical model of the hand articulation was more complex. If we consider the palm as the carrier of the hand (the 50 markers undergo a rigid motion that is computed as the optimal translation and rotation of the 11 markers glued on the back of the hand, a reference configuration of the markers being chosen with all fingers out), the movements of the wrist, the palm and the phalanges of the fingers have quite a complex nonlinear influence on the 3-D positions of the markers. These positions also reflect poorly on the underlying rotations of the joints: skin deformations induced by the muscle and skin tissues produce very large variations of the distances between markers glued on the same phalange.

The model of hand articulation was built in four steps.

(1) An estimation of the hand movement using the hypothesis of a rigid motion of markers placed on the back of the hand in corpus 3. A principal component analysis of the six parameters of this hand motion was then performed and the *nmH* (stands for *number of free movements of the hand*) first components were retained as control parameters for the hand motion.
(2) For all frames from corpus 1 and 3, where the 50 markers were all visible, all possible angles between each hand segment and the back of the hand as well as between successive phalanges were computed (rotation, twisting, spreading,…).
(3) A principal component analysis of these *nH* angles was then performed and the *naH* (stands for *number of free articulatory movements of the hand shape*) first components were retained as control parameters for the hand shaping.
(4) The sin( ) and cos( ) of these *predicted* values were computed and a linear regression between these $2*nH+1$ values (see vector P below) and the 3-D coordinates of the hand markers (see matrix *Xang* below) was performed (subtracting the inverse rigid motion of the full hand data).

Step (4) made the hypothesis that the displacement induced by a pure joint rotation may produce an elliptic movement on the skin surface (together with a scaling factor).

The final algorithm for computing the 3-D positions *P3DH* of the 50 hand markers for a given frame is as follows:

```
mvt = mean_mH+pmH*eigv_mH;

ang = mean_A+paH*eigv_A;

P = [1 cos(ang) sin(ang)];

P3DH = Rigid_Motion
       x(reshape(P*Xang,3,50),mvt);
```

where *mvt* is the movement of the back of the hand controlled by the *nmH* parameters *pmH* and *ang* is the set of angles controlled by the *naH* parameters *paH*.

C. Modeling results

After pruning corpus 1 and 3, the training data for constructing the hand shape model consisted of 8446 frames. After pruning corpora 2 and 3, the training data for facial movements consisted of 4938 frames. The number of elementary angles nH is equal to 23. Figure 3 shows the reduction of variance obtained by keeping $naH=12$ hand shape parameters and $naF=7$ face parameters. Figure 4 shows an example of a raw motion capture frame and the predicted hand and face shapes.

We retained $nmF=5$ and $nmH=5$ parameters for the head and hand movements.

Using the first 68 utterances of corpus 3 as training data (68641 frames) and a joint estimation of hand motion and hand shaping (resp., head motion and facial movements), the resulting average absolute modeling error for the position of the visible markers was 1.2 mm for the hand and 1 mm for the face (see Fig. 5). Regularization of the test data (the next 170 utterances) by the hand and face models do not lead to a substantial increase of the mean reconstruction error.

IV. FURTHER DATA ANALYSIS

Further data analysis was performed in order to verify that the cuer had realized the recommended hand shapes and

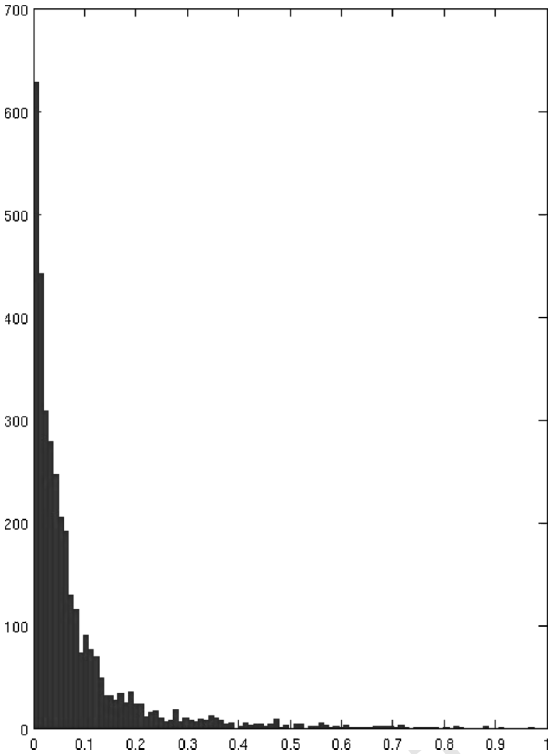


FIG. 6. Histogram of the percentage of distance accomplished by the head of our speech cuer for producing hand/face constrictions.

hand positions with the consonants and vowels effectively. In the following, all available frames were considered. Movements and articulations of hand and face were regularized and reconstructed using the hand and face models described above.

A. The constriction model

Globally the FCS functions as a constriction model: with a certain shape of the final effector (i.e., the hand), a constriction—most of the time a full contact, i.e., an occlusion—is made between the hand and the face. The *place of constriction* is determined by the vowel and the shape of the effector is determined by the consonant. Contrary to vocal tract constrictions where the walls are almost rigid—with the exception of the lips and the velum—cued speech constrictions concern two movable body segments, i.e., the head and the hand. If head movements were known to contribute to the encoding of the linguistic structure of the utterance and signals cognitive activities of the speaker, the head movements here also participated in the realization of hand–face constrictions: during speech, the head and the hand both move toward each other (see Fig. 6) and the chin or the throat to the hand according to the required hand placements, i.e., finger/head constrictions. We computed the relative displacements of the head and hand for producing hand/face constrictions by considering the maximum distance between the hand and the face in the interval between successive target hand shapes and placements. Figure 6 shows the histogram of the percentage of this distance accomplished by the head. The mean contribution is 7.7%.

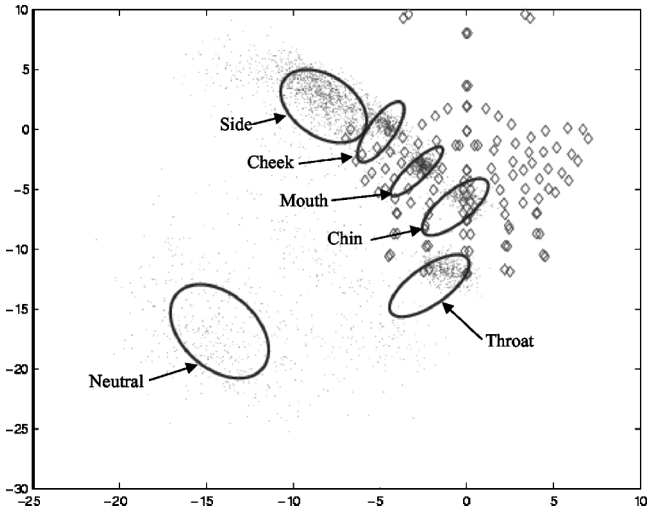


FIG. 7. Data and dispersion ellipsis of the position of finger tip of the longest finger for each targeted hand placements. Please note the main orientation of the four dispersion ellipsis for throat, chin, mouth and cheek. As expected, neutral and side hand placements exhibit the larger dispersion ellipsis (with the main axis perpendicular to the others).

B. Recognizing hand shapes and consonants

According to our previous experience (Attina *et al.*, 2002), the maximal extension/retraction of the fingers—i.e., the hand shape target—was roughly synchronized with the acoustic onset of the consonant (and of the vowel in case of a vowel not preceded by a consonant). We thus selected target frames in the vicinity of this relevant acoustic event and labeled them with the appropriate key value, i.e., a number between 0 and 8 (see Fig. 1): 0 was dedicated to the rest

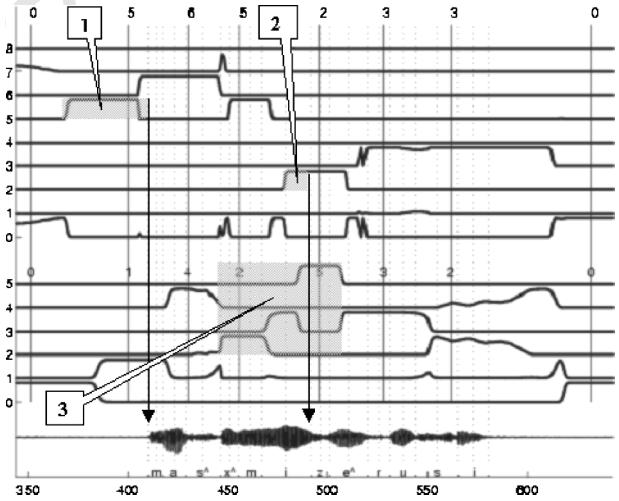


FIG. 8. Recognition of the hand shapes (top) and hand placements (bottom) by simple Gaussian models. The vertical lines show hand targets together with the required hand shapes and placements. Note that intermediate models may be triggered by a movement between hand targets. See, for example, the transition (zone 3 enlightened with translucent gray) between the hand placements 2 and 5: the model for hand placements 3 is naturally triggered (the hand goes nears the chin while moving from the mouth to the throat). Consonants are often cued well in advance of their acoustic onset (see zones 1 and 2): for example, the hand shape 5 for the first consonant [m] is deployed 50/.12=416 ms before its acoustic onset. Similarly, the hand shape 2 that signals the fourth consonant [z] is deployed 15/.12=125 ms before its acoustic onset.

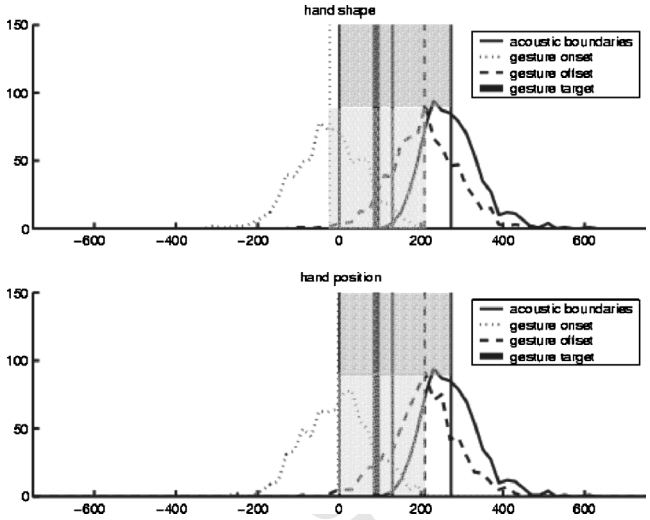


FIG. 9. Phasing gestures with reference to the different acoustic segments they are cueing. Distributions of absolute time difference of different events with reference to the acoustic onset of the segment. Both hand shapes and hand placements start well before the acoustic onset of the speech segment they are supposed to disambiguate.

position chosen by the cuer with a closed knuckle. These target frames were carefully chosen by plotting the values of seven parameters against time.

(i) For each finger, the absolute distance between the flesh point of the first phalange closest to the palm and that closest to the finger tip: a maximal value indicated an extension whereas a minimal value cued a retraction.
(ii) The absolute distance between the tips of the index and middle finger in order to differentiate between hand shapes 2 vs 8.
(iii) The absolute distance between the tip of the thumb and the palm in order to differentiate between hand shapes 1 vs 6 and 2 vs 7.

The 4114 hand shapes were identified and labeled. The seven characteristic parameters associated with these target hand shapes were then collected and simple Gaussian models were estimated for each hand shape. The *a posteriori* probability for each frame belonging to each of the eight hand shape models can then be estimated. We exhibit in Fig. 8 an example of the time course of these probability functions over the first utterance of the corpus together with the acoustic signal. Large anticipatory patterns revealed that the lip shape effectively acts as a *complementary information* to the hand shape (see Sec. IV A).

The recognition rate was quite high: there were only 4 omissions and 50 errors for a recognition rate of 98.78% (see the detailed confusion matrix in Table I). The errors involved mainly confusions between the coding of mid-vowels (/e/ vs /ɛ/ /o/ vs /ɔ/) and omissions of the coding of glides in complex CCCV sequences (such as /ʁ/ in /stʁɔ/).

C. Recognizing hand placements and vowels

Not only did the maximal extension/retraction of the fingers coincide most of the time with the acoustic onset of the consonant, but also the hand placement: CS provided both

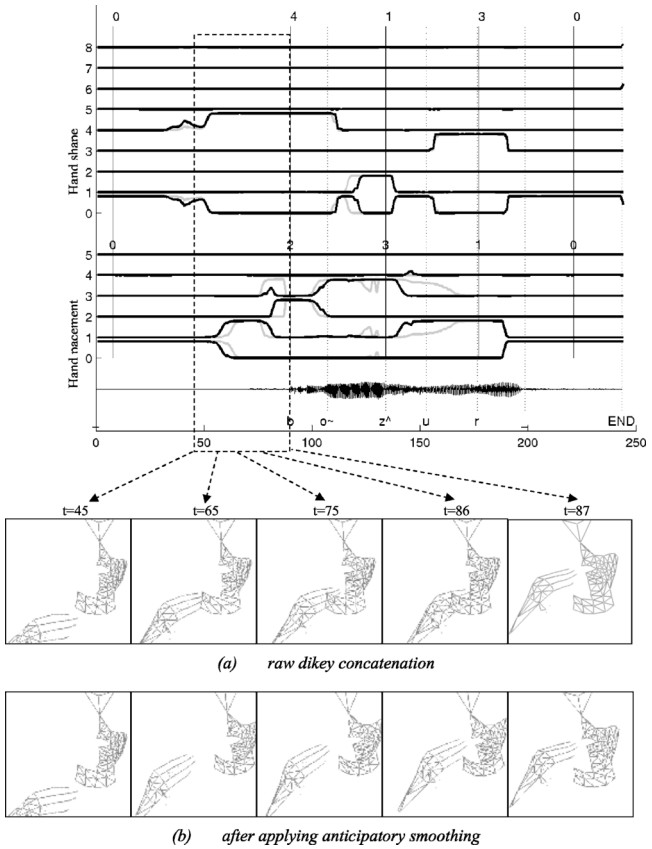


FIG. 10. Synthesis of the word "Bonjour!" ([_bɔ̃ʒuʁ_] means Hello!) by the audiovisual text-to-cued speech system. Two chronographs of the hand and face gestures generated for sequence [_bɔ̃] are shown (a) after raw concatenation and (b) after applying the smoothing procedure. The interval on the score is evidenced by a dotted rectangle. Since the dikey [00-24], is not in the dikey dictionary, the system has selected the nearest dikey in the dictionary, i.e., [00-34]. The raw dikey concatenation produces the expected movement discontinuity between frame 86 and 87 while the anticipatory smoothing procedure corrects nicely both the hand and face movements. Please check the effects of this procedure on the recognition of the hand shapes and placements (top caption): raw concatenation (data in light gray) triggers consecutively the hand placements 0, 1, 3, and 2, whereas the smoothing procedure (data in back) restores the good sequence 0, 1, and 2. Also see the increased anticipation for cueing the final [r].

the upcoming vowel and the consonant *together*, far ahead of the actual realization of the segments.

We thus added to the labels of nine hand shape targets set by the procedure described previously, the appropriate hand placement value, i.e., a number between 0 and 5 (0 for the rest position, i.e., the same as above, the cueing hand with the close knuckle being far from the face). Additional targets were also added for single vowels and start/end hand positions. Targets for single vowels were labeled with hand shape 5 while the rest position was labeled with hand placement 0.

We characterized the hand placement for these target configurations in a 3-D referential linked to the head: the 3-D position of the longest finger (index for hand shape 1 and 6 and middle finger for the others) were collected and simple Gaussian models were estimated for each hand placement.

Of the 4114 hand placements, 96.76% were identified with a total of 133 errors (Table I). There were two main sources of incorrect identifications.

TABLE I. Confusion matrices. Left: for hand shapes; right: for hand placements.

		Expected hand shapes								
		0	1	2	3	4	5	6	7	8
Recognized	0	462	1	1	4	4	15	0	0	3
	1	0	482	0	1	0	0	1	0	0
	2	1	0	419	0	0	1	0	0	0
	3	0	0	0	598	0	0	0	0	0
	4	2	0	2	0	357	0	0	0	0
	5	9	0	0	1	0	957	0	1	0
	6	2	1	0	0	0	0	545	1	0
	7	0	0	0	0	0	0	0	82	0
	8	0	0	0	0	0	0	0	0	162

		Expected hand placements					
		0	1	2	3	4	5
Recognized	0	475	27	6	5	3	8
	1	1	1647	2	2	3	0
	2	0	14	565	3	7	0
	3	0	9	6	361	0	3
	4	0	9	9	1	359	0
	5	0	1	4	8	2	574

(i) Hand placement 1 (side). This hand placement was used for a consonant followed by another consonant or a *schwa* and an undershoot of this short target occurs very often (i.e., the cuer only points to the side but does not reach it). The tendency to undershoot is clearly shown in Fig. 7.

(ii) Hand placement 0 displays a large variance (see Fig. 7) and hand placements 1 (side) and 4 (cheeks) realized too far away from the face were sometimes captured by the Gaussian model as hand placement 0.

D. Phasing speech and gestures

In the present study, the data also gathered valuable information on phasing relations between speech and hand gestures and confirmed the advance of the hand onset gesture already hypothesized by Attina *et al.* (2003a). On the basis of the gestural scores provided by the cued speech decoder presented in Sec. IV B, we analyzed the profile of hand shape and hand placement gestures with reference to the acoustic realization of the speech segment they are related to (hand shape for consonants and hand placement for vowels). The extension of a gesture was defined as the time interval where the probability of the appropriate key (shape or placement) dominates the other competing keys. We excluded from the analysis the segments that required the succession of two identical keys. A sketch of the profiles for CV, V-only, and C-only sequences is presented in Fig. 9 and Table II). These results extended the data provided by Attina et al: despite an important dispersion of the data, both the hand shape and hand placement were realized well before the acoustic onset of the speech segment they relate to. Furthermore, the hand shape and hand placement gestures were highly synchronized since they participated both in the hand/head constriction, as amplified above. We tested several hypothesizes on these phasing profiles. The conclusions are as follows.

(i) For CV segments: hand placement onsets are significantly synchronized with the acoustic onset of the segment ($p<0.05$), hand shape onsets being notably in advance. Their offsets are within the second part of the vowel ($p<0.01$). The target (labeled by hand at the center of the holding of both the hand shape and placement) remain within the consonant (a mean delay of 89 ms for an average consonantal closure of 129 ms): This result is an important one because it validates the conclusions of Attina and colleagues on a more extensive corpus, i.e., the synchronization of the hand with the beginning of the CV syllable, the duration of the hand placement until the beginning of the vowel then its move during the vowel; towards the placement of the next CV syllable.

(ii) For C-only segments, hand shape and position onsets are significantly in advance of the acoustic onset of the consonant ($p<10e-9$). Their targets are synchronized with the acoustic onset of the consonant ($p<0.05$). The offset of the hand shape is within the consonant. The offset of the hand position is synchronized with the acoustic offset of the consonant ($p<0.01$).

(iii) For V-only segments, hand shape and position onsets and targets are significantly in advance of the acoustic onset of the vowel ($p<10e-9$). Their targets are synchronized with the acoustic onset of the vowel ($p<0.05$). Their offsets are realized within the vowel.

TABLE II. Average delay (ms) between the acoustic onset of CV, C-only or V-only segments and the onset, offset, and target position of the hand shape and position gestures. Remark: only the segments whose hand shapes and placements differ from their neighbors are taken into account.

Nb	CV 1027		C 474		V 182	
	Shape	Position	Shape	Position	Shape	Position
Onset	−191	−189	−234	−386	−517	−324
Offset	174	218	24	248	62	15
Target	−24	−24	−110	−110	−177	−177

TABLE III. Number of «dikeys»: transitions between two hand targets. Left: for hand shapes; right: for hand placements.

		Next hand shape								
		0	1	2	3	4	5	6	7	8
Previous	0	0	36	12	26	11	80	69	1	3
	1	27	38	55	99	49	118	62	14	22
	2	27	44	45	65	41	95	69	7	29
	3	47	89	68	74	61	157	67	11	30
	4	28	43	38	47	23	74	75	13	20
	5	47	123	94	183	105	244	130	15	31
	6	37	83	87	71	48	138	41	17	24
	7	7	5	9	20	10	13	16	3	1
	8	18	23	14	19	13	53	17	3	5

		Next hand placement					
		0	1	2	3	4	5
Previous	0	0	70	34	27	51	56
	1	127	689	305	172	168	246
	2	32	306	77	47	39	91
	3	14	288	22	21	10	25
	4	19	142	78	46	42	47
	5	46	212	76	67	64	120

V. TOWARD AN AUDIOVISUAL TEXT-TO-CUED SPEECH SYNTHESIS SYSTEM

This corpus provided an extensive coverage of the movements implied by FCS and we have designed a first audiovisual text-to-cued speech synthesis system using concatenation of multimodal speech segments. Concatenative synthesis using a large speech database and multirepresented speech units has been largely used for acoustic synthesis (Campbell, 1997; Hunt and Black, 1996) and more recently for facial animation (Minnis and Breen, 1998). This system is—to our knowledge—the first system attempting to generate hand and face movements and articulations together with speech using the concatenation of gestures and acoustics. Two units will be considered below: diphones for the generation of the acoustic signal and facial movements; and dikeys[2] for the generation of head motion as well as for hand movements and articulations.

A. Coverage of the corpus: towards text-to-cued speech synthesis

This corpus was designed initially for acoustic concatenative speech synthesis. The coverage of polysounds (the part of speech comprised between successive stable allophones, i.e., similar to diphones but excluding glides as stable allophones) was quasioptimal: we collected a minimum number of two occurrences of each polysound with a small number of utterances.

Although not quite independent (see Sec. IV), hand placements, and hand shapes were almost orthogonal. The coverage of the corpus in terms of successions of hand placements and hand shapes was quite satisfactory (see Table III): no succession of hand shapes nor hand placements was missing.

A first text-to-cued speech system had been developed using these data. This system proceeds in three steps.

(i) A prosodic model (Bailly, 2004; Bailly and Holm, submitted) trained using corpus 3 computes phoneme durations from the linguistic structure of the sentences.
(ii) Sound and facial movements are handled by a first concatenative synthesis using polysounds (and diphones if necessary) as basic units.
(iii) Head movements, hand movements, and hand shaping movements are handled by a second concatenative synthesis using dikeys (see below) as basic units.

A key (hand and head gesture) will be referenced in the following by two numbers representing the hand placement and hand shape. For example, the key 24 (hand placement 2 together with hand shape 4) will be selected for cueing the CV sequence [bõ]. The so-called dikeys are part of movements comprised between two successive keys. For example, the dikey [00-24] stores the hand and head movements from the rest position 00 toward the key 24. Once selected, the onsets of these dikeys were further aligned with the acoustic C mid-point for full CV realizations, vocalic onsets for "isolated" vowels (not immediately preceded by a consonant) and consonantal onsets for "isolated" consonants (not immediately followed by a vowel). This phasing relation is in accordance with the data presented in Sec. IV D. If the full dikey does not exist, replacement dikeys are found by replacing the second hand placement of the dikey by the closest one that does exist in the dikey dictionary. The proper dikey will then be realized through the application of an *anticipatory* smoothing procedure (Bailly, 2002) that considered the onset of each dikey as the intended target of the first key: a linear interpolation of the parameters for the hand model is gradually applied within the preceding dikey in order that its final target coincide with the onset of the current dikey.

An example of a sequence generated by the proposed system is shown in Fig. 10. The sequence results from the concatenation and smoothing of four dikeys selected from four different utterances. Automatic recognition of the hand shapes and placements (similar to Fig. 8) is provided in order to demonstrate the ability of the system to generate the appropriate gestures. This two-step procedure generated acceptable synthetic cued speech. It, however, considers the head movements to be entirely part of the realization of hand-face constrictions and uses, for now a crude approximation of the speech/gesture coordination (although being a very satisfactory first-order approximation as suggested by Attina *et al.*, 2002).

B. From gestures to appearance

The text-to-cued speech synthesis system sketched above delivered trajectories of a few flesh points placed on

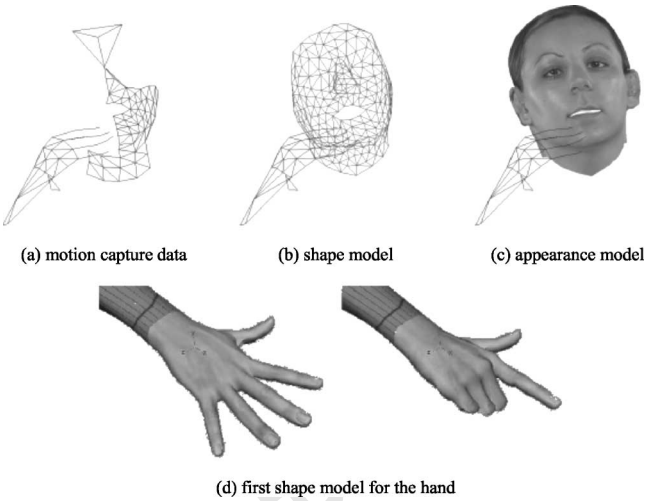


FIG. 11. Driving a complete shape and video-realistic appearance model of the cuer from the parameters of the face model. (a) Initial facial flesh points; (b) high-definition mesh; (c) textured mesh. High-definition meshes and video-realistic textures for the hand [see (d)], the teeth and the inner mouth are currently under development.

the surface of the right hand and face. We plan to evaluate the benefits brought by this system in speech understanding using the point-light paradigm we already used for face only (Bailly *et al.*, 2002; Odisio and Bailly, 2004).

We are currently interfacing this trajectory planning with a detailed shape and appearance model of the face and hand of the original speaker. High definition models of these organs—comprising several hundreds of vertices and polygons—are first mapped onto the existing face and hand parameter space. A further appearance model using video-realistic textures is then added (Elisei *et al.*, 2001; Bailly *et al.*, 2003). Figure 11 illustrates our ongoing effort toward the animation of a videorealistic virtual cuer. Applying the same procedure to a high definition model of the hand is currently under study.

C. Comments

An important challenge of the phonetic description of FCS is clearly understanding the constraints acting on the multisegment gestural planning: several segments are, in fact, recruited in addition to the usual speech articulators (jaw, lips, tongue, larynx,…). Head and hand movements should be appropriately phased in order to ease lip reading. We suggest here that this planning be done in terms of constrictions made while recruiting specific hand configurations. This planning ensures that cued speech information will be delivered well in advance of the lip reading information. This gestural score has interesting similarities with data on synchronization between deictic pointing and speech that evidences also an important anticipatory pointing gesture in relation with acoustic onset (as large as 300 ms in Castiello *et al.*, 1991). For an optimal processing of the message, the "*where*" component of the multimodal deictic gesture should precede the "*who*" component.

Movement execution recruits segments in an optimal manner, but all segments actually participate in the realization of the series of constrictions. A major challenge for movement analysis and generation will be to separate out prosodic—or literally suprasegmental—movements of the head from movements of the head contributing to the correct decoding/encoding of the speech segments.

Another important issue will be to understand what the influence of FCS is, on the temporal organization of the speech gestures and more specifically on the rhythmical organization of the speech stream.

VI. CONCLUSIONS AND PERSPECTIVES

The immense benefits of Cued Speech in terms of giving access to language structure and speech comprehension to deaf people should be grounded in a deep understanding of its implementation by actual speakers. Although precise qualitative guidelines have been specified by Cornett, the FCS is an evolving system whose phonetic structure is constantly enriched by cuers.

We analyzed here the live recordings of the hand, face and head gestures with reference to the phonetic structure of the speech sounds produced by a user of FCS. When compared to the expected hand shapes and hand placement targets positioned by the hand, the automatic cued speech recognizer operating on hand gestural scores identified respectively 98.14% and 95.52% of the hand shapes and placements. The errors could clearly be interpreted in terms of undershoot or true phonetic confusions, mostly involving confusions between vowel apertures or consonant devoicing/voicing.

The study of the phasing relation between these gestural scores and the phonetic structure of the sound produced confirmed the empirical rules used by Duchnowski *et al.* (2000) for automatically superimposing virtual hand shapes on a prerecorded video of a speaking person and the motion capture data analyzed by Attina *et al.* (2002; 2003a; 2004): the hand movements provided phonetic cues for the decoding of the incoming speech well before any other acoustic or facial cues; typically more than 200 ms before the acoustic onset of the sound the gesture related to. Further perception experiments involving gating experiments and reaction times will be required to test if this cued speech advantage is actually used by observers and to test the sensitivity of their performance when this anticipatory coarticulation is altered.

The observation of cuers in action is thus a prerequisite for developing technologies that will assist deaf people in learning FCS. The database recorded, analyzed, and characterized here is currently exploited within a multimodal text-to-FCS system that will supplement or replace on-demand subtitling by a virtual FCS cuer for TV broadcasting or home entertainment. With the ARTUS project, ICP and Attitude Studio collaborate with academic and industrial partners in order to provide the French–German TV channel ARTE with the possibility of broadcasting programs dubbed with virtual CS. The movements of which are computed from existing subtitling or captured life on a FCS interpreter, watermarked within the video and acoustic channels and rendered locally by the TV set. The low transmission rate of CS as required

by watermarking should also benefit from a better understanding of the kinematics of the different segments involved in the production of CS.

ACKNOWLEDGMENTS

Many thanks to Yasmine Badsi, our CS speaker for having accepted the recording constraints. We thank Christophe Corréani, Xavier Jacolot, Jeremy Meunier, Frédéric Vandenbergh, and Franck Vayssettes for the processing of the raw motion capture data. We also acknowledge Virginie Attina for providing her cued speech expertise when needed. We thank Siyi Wang for helping us to correct the English. This work is financed by the RNRT ARTUS. We acknowledge three anonymous reviewers for thoughtful remarks on the earlier versions of this paper.

[1]Throughout this article, the use of the term cuer refers to a user of cued speech.

[2]A dikey is defined as the part of movements comprised between two successive keys (see Sec. V A).

Attina, V., Beautemps, D., and Cathiard, M.-A. (**2002**). "Coordination of hand and orofacial movements for CV sequences in French Cued Speech," *International Conference on Speech and Language Processing*, Boulder, pp. 1945–1948.

Attina, V., Beautemps, D., and Cathiard, M.-A. (**2003a**). "Temporal organization of French Cued Speech production," *International Conference of Phonetic Sciences*, Barcelona, Spain.

Attina, V., Beautemps, D., Cathiard, M.-A., and Odisio, M. (**2003b**). "Towards an audiovisual synthesizer for Cued Speech: rules for CV French syllables," *Auditory–Visual Speech Processing*, St Jorioz, France, pp. 227–232.

Attina, V., Beautemps, D., Cathiard, M.-A., and Odisio, M. (**2004**). "A pilot study of temporal organization in Cued Speech production of French syllables: rules for a Cued Speech synthesizer," Speech Commun. **44**, 197–214.

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

Bailly, G., and Badin, P. (**2002**). "Seeing tongue movements from outside," *International Conference on Speech and Language Processing*, Boulder, Colorado, pp. 1913–1916.

Bailly, G., Gibert, G., and Odisio, M. (**2002**). "Evaluation of movement generation systems using the point-light technique," *IEEE Workshop on Speech Synthesis*, Santa Monica, CA, pp. 27–30.

Bailly, G., Holm, B., and Aubergé, V. (**2004**). "A trainable prosodic model: learning the contours implementing communicative functions within a superpositional model of intonation," *International Conference on Speech and Language Processing*, Jeju, Korea, pp. 1425–1428.

Bailly, G., and Holm, B. (to be published). "SFC: a trainable prosodic model," Speech Commun..

Bailly, G., Bérar, M., Elisei, F., and Odisio, M. (**2003**). "Audiovisual speech synthesis," International Journal of Speech Technology **6**(4) 331–346.

Bernstein, L. E., Demorest, M. E., and Tucker, P. E. (**2000**). "Speech perception without hearing," Percept. Psychophys. **62** 233–252.

Campbell, N. (**1997**). "Computing prosody: Computational models for processing spontaneous speech," *Synthesizing Spontaneous Speech*, edited by Y. Sagisaka, N. Campbell, and N. Higuchi (Springer-Verlag, Berlin), pp. 165–186.

Castiello, U., Paulignan, Y., and Jeannerod, M. (**1991**). "Temporal dissociation of motor responses and subjective awareness," Brain **114**, 2639–2655.

Cornett, R. O. (**1967**). "Cued Speech," Am. Ann. Deaf **112**, 3–13.

Cornett, R. O. (**1988**). "Cued Speech, manual complement to lipreading, for visual reception of spoken language. Principles, practice and prospects for automation," Acta Otorhinolaryngol. Belg. **42**, 375–384.

Cornett, R. O., and Daisey, M. E. (**1992**). *The Cued Speech Resource Book for Parents of Deaf Children*. Raleigh, NC, The National Cued Speech Association, Inc.

Duchnowski, P., Lum, D. S., Krause, J. C., Sexton, M. G., Bratakos, M. S., and Braida, L. D. (**2000**). "Development of speechreading supplements based on automatic speech recognition," IEEE Trans. Biomed. Eng. **47**(4): 487–496.

Elisei, F., Odisio, M., Bailly, G., and Badin, P. (**2001**). "Creating and controlling video-realistic talking heads," *Auditory-Visual Speech Processing Workshop*, Scheelsminde, Denmark, pp. 90–97.

Engwall, O., and Beskow, J. (**2003**). "Resynthesis of 3D tongue movements from facial data," *EuroSpeech*, Geneva.

Hunt, A. J., and Black, A. W. (**1996**). "Unit selection in a concatenative speech synthesis system using a large speech database," *International Conference on Acoustics, Speech and Signal Processing*, Atlanta, GA, pp. 373–376.

Jiang, J., Alwan, A., Bernstein, L., Keating, P., and Auer, E. (**2000**). "On the Correlation between facial movements, tongue movements and speech acoustics," *International Conference on Speech and Language Processing*, Beijing, China, pp. 42–45.

Kuratate, T., Munhall, K. G., Rubin, P. E., Vatikioti-Bateson, E., and Yehia, H. (**1999**). "Audio-visual synthesis of talking faces from speech production correlates," *EuroSpeech*, pp. 1279–1282.

Leybaert, J. (**2000**). "Phonology acquired through the eyes and spelling in deaf children," J. Exp. Child Psychol. **75**, 291–318.

Leybaert, J. (**2003**). "The role of Cued Speech in language processing by deaf children: An overview," *Auditory-Visual Speech Processing*, St Jorioz, France, pp. 179–186.

Minnis, S., and Breen, A. P. (**1998**). "Modeling visual coarticulation in synthetic talking heads using a lip motion unit inventory with concatenative synthesis," *International Conference on Speech and Language Processing*, Beijing, China, pp. 759–762.

Nicholls, G., and Ling, D. (**1982**). "Cued Speech and the reception of spoken language," J. Speech Hear. Res. **25**, 262–269.

Odisio, M., and Bailly, G. (**2004**). "Shape and appearance models of talking faces for model-based tracking," Speech Commun. **44**, 63–82.

Owens, E., and Blazek, B. (**1985**). "Visemes observed by hearing-impaired and normal-hearing adult viewers," J. Speech Hear. Res. **28**, 381–393.

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

Uchanski, R., Delhorne, L., Dix, A., Braida, L., Reed, C., and Durlach, N. (**1994**). "Automatic speech recognition to aid the hearing impaired: Prospects for the automatic generation of cued speech," J. Rehabil. Res. Dev. **31**, 20–41.

Yehia, H. C., Rubin, P. E., and Vatikiotis-Bateson, E. (**1998**). "Quantitative association of vocal-tract and facial behavior," Speech Commun. **26**, 23–43.

PROOF COPY 025508JAS