AUDIOVISUAL TEXT-TO-CUED SPEECH SYNTHESIS

Guillaume Gibert, Gérard Bailly, Frédéric Elisei

Institut de la Communication Parlée (ICP), UMR CNRS 5009, INPG/U3
46, av. Félix Viallet – 38031 Grenoble France
{gibert, bailly, elisei}@icp.inpg.fr

ABSTRACT

We present here our efforts for implementing a system able to synthesize French Manual Cued Speech (FMCS). We recorded and analyzed the 3D trajectories of 50 hand and 63 facial flesh points during the production of 238 utterances carefully designed for covering all possible diphones of the French language. Linear and non linear statistical models of the hand and face deformations and postures have been developed using both separate and joint corpora. We create 2 separate dictionaries, one containing diphones and another one containing "dikeys". Using these 2 dictionaries, we implement a complete text-to-cued speech synthesis system by concatenation of diphones and dikeys.

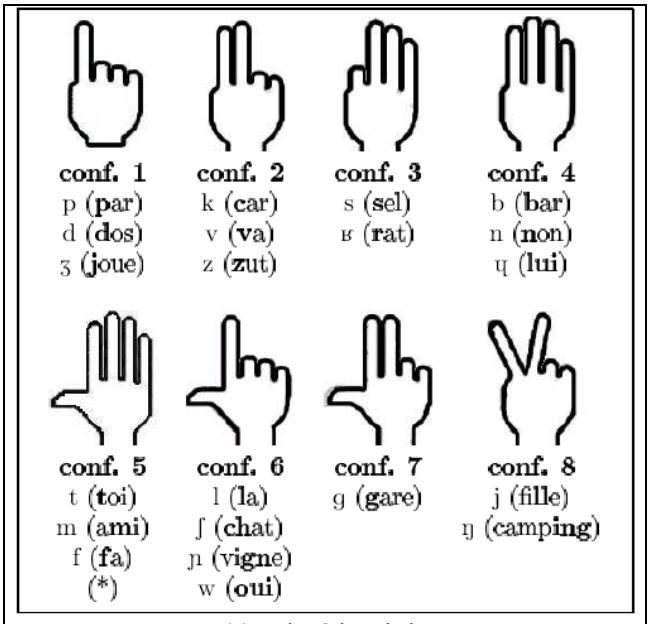
1. INTRODUCTION

Speech articulation has clear visible consequences. If the movements of the jaw, the lips and the cheeks are immediately visible, 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 weakly correlated with visible movements (R ~ 0.7) [11, 15] and this correlation is insufficient for recovering essential phonetic cues such as place of articulation [2, 9]. Listeners with hearing loss and orally educated typically rely heavily on speech reading based on lips and face visual information. However lip-reading alone is not sufficient due to the lack of information on the place of tongue articulation, the mode of articulation (nasality or voicing) and to the similarity of the lip shapes of some speech units (so called labial sosies as [u] vs. [y]). Indeed, even the best speech readers do not identify more than 50 percent of phonemes in nonsense syllables [13] or in words or sentences [5].

Manual Cued Speech (MCS) was designed to complement speechreading. Developed by Cornett [6] and adapted to more than 50 languages [7], this system is based on the association of speech articulation with cues formed by the hand.

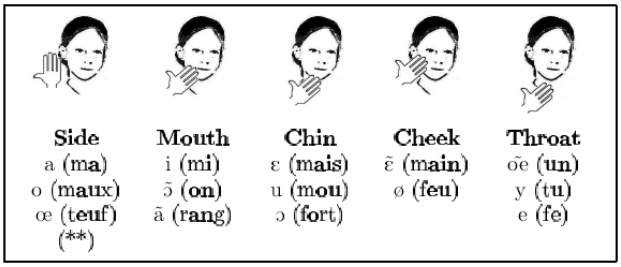
While uttering, the speaker uses one of his hand to point out specific positions on the face (indicating a subset of vowels) with a hand shape (indicating a subset of consonants as shown in Figure 1. A large amount of work has been devoted to MCS perception but few works have provided insights in the MCS production.

We describe here a series of experiments for gathering data and characterizing the hand and face movements of a FMCS speaker in order to implement a multi-modal text-to-speech synthesizer.



(a) The 8 hand shapes.

(*): Conf. 5 is also used for a vowel not preceded by a consonant.



(b) The 5 hand placements.

(**): Side is also used for a consonant followed by another consonant or a schwa.

Figure 1: French Cued Speech system.

2. MOTION CAPTURE DATA

We recorded the 3D positions of 113 markers glued on the hands and face of the subject (see Figure 2) using a Vicon© motion capture system with 12 cameras. The basic system delivers the 3D positions of candidate markers at 120Hz.

Two different settings of the cameras enabled us to record three corpora:

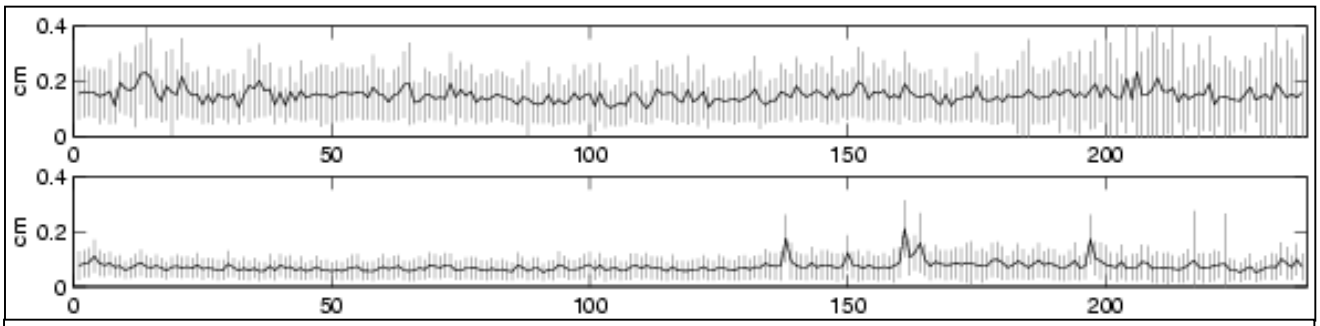


Figure 3: Mean and standard deviation of the main reconstruction error for each sentence processed by the hand (top) and the face (bottom) models (learnt with all the sentences)**.**

- a corpus of handshapes transitions produced in free space: the cuer produces all possible transitions between the eight consonantal hand shapes.
- a corpus of visemes with no hand shape associated. It consists in the production of all isolated French vowels and all consonants in symmetrical context VCV, where V is one the extreme vowels [a], [i] and [u]. This corpus is similar to the one usually used at ICP for cloning speakers [1].
- a corpus of 238 sentences pronounced with cueing the FMCS.

Corpora 1 and 2 are used to build statistical models of the hand and face movements separately.

The models are then used to recover missing data in the corpus 3: when cueing the FMCS, the face obviously hides parts of the hands and vice versa.

Figure 2: Position of the reflexive markers

3. MODELS

The scientific motivation of building statistical models from raw motion capture concern the study of FMCS: if the positions of markers are always accessible and reliable, the kinematics of the articulation, of the fingertips and fingers/face constrictions offer an unique way for studying the production of FMCS and the laws governing the coordination between acoustics, face and hand movements during cued speech production.

3.1. Face

The basic methodology developed at ICP for cloning facial articulation consists of an iterative linear analysis (Revéret, Bailly et al. 2000; Badin, Bailly et al. 2002) using the first principal component of different subsets of fleshpoints: we thus subtract iteratively the contribution of the jaw rotation, the lips rounding/spreading gesture, the proper vertical movements of upper and lower lips, of the lip corners as well as the movement of the throat to the residual data obtained by iteratively subtracting their contributions to the original motion capture data.

This basic methodology is normally applied to quasi-static heads. Since the movements of the head are free in the corpora 2 and 3, we need to solve the problem of the repartition of the variance of the positions of the 18 markers placed on the throat between head and face movements. This problem is solved in three steps:

- Estimation of the head movement using the hypothesis of a rigid motion of markers placed on the ears, nose and forehead. A principal component analysis of the 6 parameters of the rototranslation extracted for corpus 3 is then performed and the *nmF* first components are retained as control parameters for the head motion.
- Facial motion cloning using the inverse rigid motion of the full data. Only *naF* components are retained as control parameters for the facial motion.
- Throat movements are considered to be equal to head movements weighted by factors less than one. A joint optimization of these weights and the directions of *nmF* facial deformations is then performed keeping the same values for the *nmF* and *naF* predictors.

These operations are performed using facial data from corpus 2 and 3 with all markers visible.

A simple vector quantization guarantying a minimum 3D distance between selected training frames (equal here to 2mm) is performed before modeling. This pruning step provides statistical models with conditioned data.

The final algorithm for computing the 3D positions of *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 TO 63
  M = mvt.*wmF(:,i);
  P3DF(:,i) = Rigid_Motion(P3D(:,i),M);
END
```

where *mvt* are the head movements controlled by the *nmF* 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 3D positions of the markers without head movements controlled by *naF* parameters *paF*.

3.2. Hand

Building a statistical model of the hand deformations is more complex. If we consider the forearm as being the carrier of the hand (the 50 markers undergo a rigid motion that will be considered as the forearm motion), the movements of the wrist, the palm and the phalanges of the fingers have quite complex non linear influence on the 3D positions of the markers. These positions reflect also poorly the underlying rotations of the joints: skin deformation 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 deformations is built in four steps:

- Estimation of the hand movement using the hypothesis of a rigid motion of markers placed on the forearm in corpus 1. A principal component analysis of the 6 parameters of this hand motion is then performed and the nmH first components are retained as control parameters for the hand motion.
- All possible angles between each hand segment and the forearm as well between successive phalanges (using the inverse rigid motion of the full hand data) are computed (rotation, twisting, spreading).
- A principal component analysis of these angles is then performed and the naH first components are retained as control parameters for the hand shaping.
- We then computed the sin() and cos() of these predicted values and perform a linear regression between these *2*naH+1* values (see vector *P* below) and the 3D coordinates of the hand markers.

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

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

```
mvt = mean_mH + pmH*eigv_mH;
ang = mean_A + paH*eigv_A;
P = [1 cos(ang) sin(ang)];
P3DH = Rigid_Motion(reshape(P*Xang,3,50),mvt);
```

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

3.3. Modeling Results

Using the corpus 1, the training data for hand shapes consists of 8446 frames. Using corpus 2 and 3, the training data for facial movements consists of 4938 frames. We retain *naH* = 12 hand shape parameters and *naF* = 7 face parameters. Using the first 68 utterances of the corpus 3 as training data (68641 frames) and a joint estimation of hand motion and hand shapes (resp. head motion and facial movements), the resulting average absolute modeling error for the position of the visible markers is equal to 2mm for the hand and 1mm for the face (see Figure 3). Inversion of the test data (the next 114 utterances) by the hand and face models do not lead to a substantial increase of the mean reconstruction error.

4. DATA ANALYSIS: PHASING

Before implementing a multimodal text-to-speech, we need to characterize the phasing rules between hand movements and speech.

Hand shape. We selected target frames in the vicinity of the relevant acoustic event and labeled them with the appropriate key value, i.e. a number between 0 and 8: 0 is dedicated to the rest position chosen by the cuer with a closed knuckle. These target frames are carefully chosen by plotting the values of 7 parameters against time:

- For each finger, the absolute distance between the fleshpoints of the first phalange closest to the palm and that closest to the finger tip: a maximal value indicates an extension whereas a minimal value cues a retraction.
- The absolute distance between the tips of the index and middle finger in order to differentiate between hand shapes 2 versus 8.
- The absolute distance between the tip of the thumb and the palm in order to differentiate between hand shapes 1 versus 6 and 2 versus 7.

4114 hand shapes were identified and labeled. The 7 characteristic parameters associated with these target hand shapes are then collected and simple Gaussian models are estimated for each hand shape. The a posteriori probability for each frame to belong to each of the 9 hand shape model can then be estimated.

Hand placement. We thus added to the labels of 9 hand shape targets - set by the procedure described above - a key value for the hand placement (between 0 and 5: 0 corresponds to the rest position). Hand shape and placement targets were added for single vowels and labeled with hand shape 5 while the rest position (closed knuckle far from the face) was labeled with hand placement 0.

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

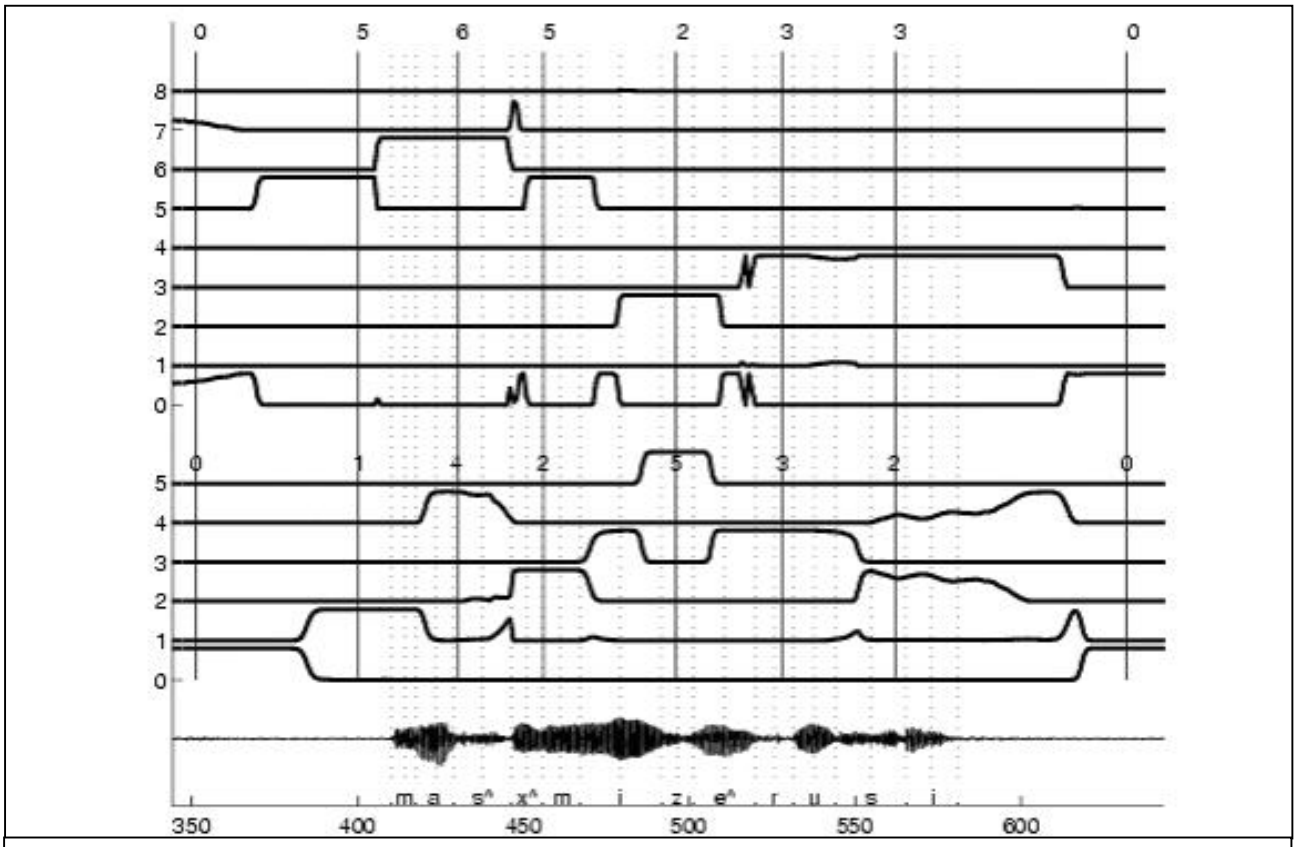


Figure 4. Recognition of the hand shape (top) and the hand placement (bottom) by simple Gaussian models.

Gestural scores. We exhibit in Figure 4 an example of the time course of these probability functions over the first utterance of the corpus together with the acoustic signal.

We analyzed the profile of hand shape and hand placement gestures. We verify manually the pre-segmentation done using our MOTHER OPENGL© animation software [14]. These gestures are then further characterized in 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 is 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.

Phasing profiles. A sketch of the profiles for CV sequences is presented in Figure 5: for a full CV realization, the hand movement (shape and position) starts quasi-synchronously with the vocalic onset, the target is reached in the middle of the consonant.

For "isolated" vowels (realized with hand shape 5) and "isolated" consonants vowels (realized with hand placement 1) , the target is reached quasi-synchronously with the vocalic onset. Furthermore the hand shape and hand placement gestures are highly synchronized since they participate both to the hand/head constriction as amplified above.

5. TOWARDS A MULTI-MODAL SPEECH SYNTHESIS SYSTEM

This corpus provides an extensive coverage of the movements implied by FMCS and we have designed a first audiovisual text-to-cued speech synthesis system using concatenation of multimodal speech segments. If concatenative synthesis using a large speech database and multi-represented speech units is largely used for acoustic synthesis [10] and more recently for facial animation [12], this system is to our knowledge the first system attempting to generate hand and face movements and deformations together with speech using the concatenation of gestural and acoustic units. Two units will be considered below: diphones for the generation of the acoustic signal and facial movements; and dikeys for the generation of head and hand movements.

5.1. Selection and concatenation

This corpus was designed initially for acoustic concatenative speech synthesis. The coverage of polysounds (part of speech comprised between successive stable allophones, i.e. similar to diphones but excluding glides as stable allophones) is quasi-optimal: we collect a minimum number of 2 samples of each polysound with a small number of utterances. Although not quite independent (see previous section), hand placements and hand shapes are almost orthogonal. The coverage of the corpus in terms of successions of hand placements and hand shapes is quite satisfactory: no succession of hand shapes nor

hand placement is missing. A first text-to-cued speech system has been developed using these data. This system proceeds in two steps:

1. the sound and facial movements are handled by a first concatenative synthesis using polysounds (and diphones if necessary) as basic units.
2. the head movements, the hand movements and the hand shaping movements are handled by a second concatenative synthesis using "dikeys" as basic units.

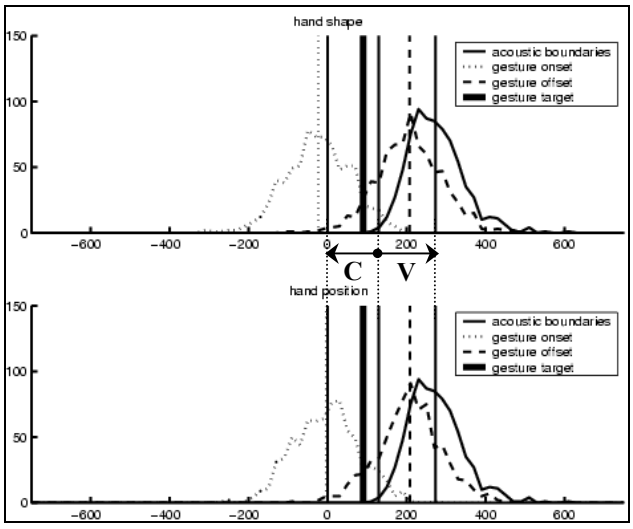


Figure 5: Phasing gestures with reference to the onsets of all CV acoustic segments cued. Associated LPC gestures are initiated almost in phase with C acoustic onsets while the gesture targets fall within the initial consonants.

Sound and facial movements The first system consists in a diphone-based audiovisual concatenation system. Diphones are multi-represented: candidate diphones are selected using a standard dynamic programming technique. The selection cost is proportional to the RMS distance between desired phonemic durations and the durations of the concatenated diphones. The concatenation cost is proportional to the RMS distance between the Line Spectrum Pairs (available in the characterization of the audio signals: we use a TDPSOLA technique for the audio manipulation) across each boundary and does not take - for now - into account any articulatory distance. Intra-diphone articulatory trajectories are warped synchronously with acoustic frames. We add an articulatory smoothing procedure to compensate jumps observed at the inter-diphone boundaries: it computes a linear interpolation of the observed jump during the previous diphone.

Hand and head movements

The second system will concatenate dikeys: a key (hand and head gesture) will be referenced in the following by two numbers figuring 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]. Analog to diphones for hand movements, the so-called dikeys are part of movements comprised between two successive keys (e.g. the dikey [00-24] stores the hand and head movements from the rest position 00 towards the key 24).

Dikeys are multi-represented: candidate are selected using a standard dynamic programming technique. The selection cost is proportional to the RMS distance between desired dikeys durations and the durations of the concatenated dikeys. The concatenation cost is proportional to the RMS distance between the articulatory parameters (weighted by the variance of the movement explained) across each boundary. Once selected, these dikeys are further aligned with the middle of C 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 quite in accordance with the data presented in §4. If the full dikey does not exist, we seek for alternative dikeys by replacing the second hand placement of the dikey by the closest one that do exist in the dikey dictionary. The proper dikey will be even realized since an anticipatory smoothing procedure (as for the facial movements) is applied that considers the onset of each dikey as the intended target: a linear interpolation of the parameters for the hand model gradually applied within each dikey copes thus easily with a small (or even larger) change of the final target imposed by the onset target frame of the next concatenated dikey.

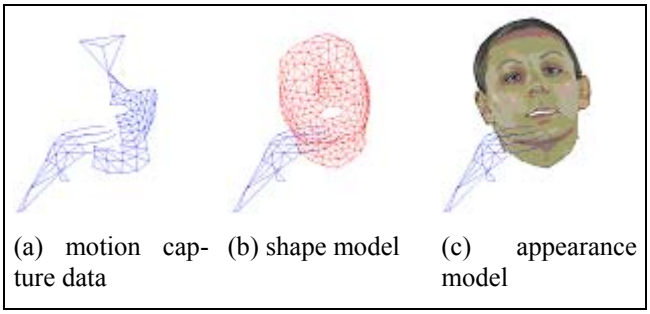
Comments. We checked that this two-step procedure generates synthetic cued speech movements well identified by the gesture recognizer (see §4). Further perceptual tests will be conducted in the future. We plan also to overcome some limitations of the current approach that considers notably the head movements to be entirely part of the realization of hand-face constrictions – obviously including a suprasegmental component that should be extracted and modeled separately - and uses for now a crude approximation of the speech/gesture coordination – that could be stored for each dikey.

5.2. Animation

The text-to-cued speech synthesis system sketched above delivers trajectories of a few flesh points placed on 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 [3].

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 – is first mapped onto the existing face and hand parameter space. A further appearance model using video-realistic textures is then added [4, 8]. Figure 6 illustrates our ongoing effort towards the animation of a video realistic virtual cuer. Applying the same procedure to a

high definition model of the hand is currently under study.



(a) motion capture data (b) shape model (c) appearance model

Figure 6: From motion capture data to a video-realistic appearance model.

6. CONCLUSIONS & PERSPECTIVES

The observation of cuers in action is a prerequisite for developing technologies that can assist deaf people in learning the FCS. The multimodal text-to-FCS speech system developed here will supplement or replace on-demand subtitling by a virtual FCS cuer for TV broadcasting or home entertainment. Within 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 a virtual CS, the movements of which are computed from existing subtitling or captured on a FCS interpreter, watermarked within the video and acoustic channels and rendered locally by the TV. Low rate transmission 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.

Several other cued-speech-specific modules are also under development, including generation of prosody (the basic CV rhythmic structure of cuing is expected to interfere with the basic syllabic organization of speech rhythm).

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 acknowledge also Virginie Attina for providing her cued speech expertise when needed. This work was financed by the RNRT ARTUS.

REFERENCES

[1] 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.* Journal of Phonetics, **30**(3): p. 533-553.

[2] Bailly, G. and Badin, P. (2002) *Seeing tongue movements from outside*. in *International Conference on Speech and Language Processing*. Boulder - Colorado. p. 1913-1916.

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

[4] Bérar, M., Bailly, G., Chabanas, M., Elisei, F., Odisio, M., and Pahan, Y. (2003) *Towards a generic talking head*. in *6th International Seminar on Speech Production*. Sydney- Australia. p. 7-12.

[5] Bernstein, L.E., Demorest, M.E., and Tucker, P.E. (2000) *Speech perception without hearing.* Perception & Psychophysics, **62**: p. 233-252.

[6] Cornett, R.O. (1967) *Cued Speech.* American Annals of the Deaf, **112**: p. 3-13.

[7] Cornett, R.O. (1988) *Cued Speech, manual complement to lipreading, for visual reception of spoken language.* Principles, practice and prospects for automation. Acta Oto-Rhino-Laryngologica Belgica, **42**(3): p. 375-384.

[8] Elisei, F., Odisio, M., Bailly, G., and Badin, P. (2001) *Creating and controlling video-realistic talking heads*. in *Auditory-Visual Speech Processing Workshop*. Scheelsminde, Denmark. p. 90-97.

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

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

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

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

[13] Owens, E. and Blazek, B. (1985) *Visemes observed by hearing-impaired and normal-hearing adult viewers.* Journal of Speech and Hearing Research, **28**: p. 381-393.

[14] 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*. in *International Conference on Speech and Language Processing*. Beijing - China. p. 755-758.

[15] Yehia, H.C., Rubin, P.E., and Vatikiotis-Bateson, E. (1998) *Quantitative association of vocal-tract and facial behavior.* Speech Communication, **26**: p. 23-43.