AUDIOVISUAL TEXT-TO-CUED SPEECH SYNTHESIS

Guillaume Gibert[1], *Gérard Bailly*[1], *Frédéric Eliséi*[1], *Denis Beautemps*[1], *Rémi Brun*[2]

(1) Institut de la Communication Parlée UMR CNRS 5009, INPG/U3
46, av. Félix Viallet - 38031 Grenoble France
phone: +33 (0)4 76 57 45 34 - fax: +33 (0)4 76 57 47 10
email: gibert@icp.inpg.fr
(2) Attitude Studio SA, 100 Avenue du Général Leclerc, 93692 Pantin France

ABSTRACT

We present here our efforts for characterizing the 3D movements of the right hand and the face of a French female during the production of manual cued speech. We analyzed the 3D trajectories of 50 hand and 63 facial fleshpoints 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 implement a concatenative audiovisual text-to-cued speech synthesis system.

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 ($R \sim 0.7$) [17, 10] and this correlation is insufficient for recovering essential phonetic cues such as place of articulation [2, 7]. Listeners with hearing loss and orally educated typically rely heavily on speechreading based on lips and face visual information. However lipreading 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 speechreaders do not identify more than 50 percent of phonemes in nonsense syllables [14] or in words or sentences [4]. Manual Cued Speech (MCS) was designed to complement speechreading. Developed by Cornett [5] and adapted to more than 50 languages [6], 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 handshape (indicating a subset of consonants as shown in figure 1). For more details on the French MCS (FMCS) system please see http://retore.chez.tiscali.fr/LPC. Numerous studies have demonstrated the drastic increase of intelligibility provided by MCS compared to lipreading alone [13, 16] and the effective facilitation of language learning using FMCS [11]. 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 cued-speech synthetizer.

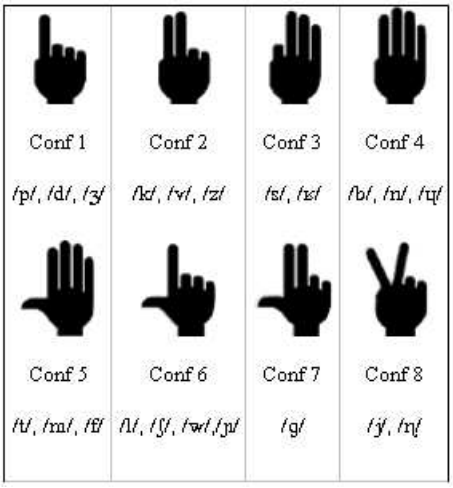


Figure 1: French Cued Speech system for consonants.

2. MOTION CAPTURE DATA

We recorded the 3D positions of 113 markers glued on the hands and face of the subject 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:

- 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 handshape 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 in 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.

3. ARTICULATORY MODELS OF THE FACE AND OF THE HAND

The scientific motivation of building statistical models from raw motion capture concern the study of FMCS: if the posi-

tions of markers are always accessible and reliable, the kinematics of the articulation, of the finger tips 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 [1, 15] 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:

$$
\begin{array}{l}
mvt = mean\_mF + pmF * eigv\_mF; \\
P3D = reshape(mean\_F + paF * eigv\_F, 3, 63); \\
\underline{\textbf{for}}\ i := 1\ \underline{\textbf{to}}\ 63 \\
\quad M = mvt. * wmF(:, i); \\
\quad P3DF(:, i) = Rigid\_Motion(P3D(:, i), M); \\
\underline{\textbf{end}}
\end{array}
$$

where *mvt* is 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

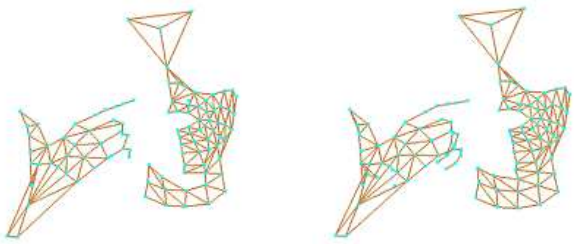
Building a statistical model of the hand deformations is more complex. If we consider the forearm as being the carrier of the hand (the 50 markers undergo a rigid motion that will be considered as the forearm motion), the movements of the wrist, the palm and the phalanges of the fingers have quite complex 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:

$$
\begin{array}{l}
mvt = mean\_mH + pmH * eigv\_mH; \\
ang = mean\_A + paH * eigv\_A; \\
P = [1 \quad cos(ang) \quad sin(ang)]; \\
P3DH = Rigid\_Motion(reshape(P * Xang, 3, 50), mvt);
\end{array}
$$

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*.



(a) incomplete motion capture data

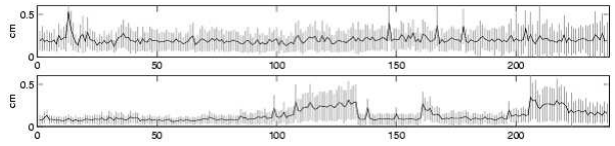
(b) reconstruction

Figure 2: 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.

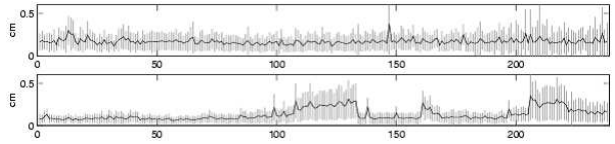
3.3 Modeling results

Using the corpus 1, the training data for handshapes consists of 8446 frames. Using corpus 2 and 3, the training data for facial movements consists of 4938 frames. We retain *naH* = 12 handshape parameters and *naF* = 7 face parameters.

Figure 2 shows an example of a raw motion capture frame and the predicted hand and face shapes. Using the first 68 utterances of the corpus 3 as training data (68641 frames) and a joint estimation of hand motion and handshapes (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). Regularization of the test data (the next 114 utterances) by the hand and face models do not lead to a substancial increase of the mean reconstruction error except for a few sentences (utterance 110 to 136) where the face error is doubled.



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

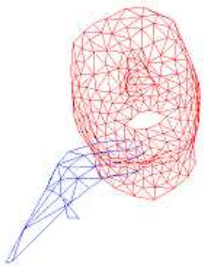


(b) learning the models with all the sentences

Figure 3: Mean and standard deviation of the mean reconstruction error for each sentence processed by the hand (top) and face (bottom) models.

(a) places of the reflexive markers

(b) motion capture data



(c) shape model

(d) appearance model

Figure 4: From motion capture data to a videorealistic virtual speech cuer. In figures (b), (c) and (d), a [$b\varepsilon$] syllabe is cued.

4. TOWARDS AN AUDIOVISUAL TEXT-TO-CUED 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 [9] 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.

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

Concatenating segments

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 minimal number of utterances. Although not quite independent, 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 : all succession of hand shapes and hand position are present. A first text-to-cued speech system has been developed using these data. This system proceeds in two steps:

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

Phasing speech and hand gesture

Once selected these dikeys are further aligned with the middle of the consonnant for a full CV realizations, vocalic onsets for ”isolated” vowels and consonantal onsets for ”isolated” consonants (this phasing relations are deduced from a preliminary data analysis [8]). If the full dikey does not exist, we seek for replacement 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 still realized because an anticipatory smoothing procedure [3] is applied that consider the onset of each dikey as the intended target: a linear interpolation of hand placement applied grad-

ually 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. This two-step procedure generates quite acceptable synthetic cued speech. It however considers the head movements to be entirely part of the realization of hand-face constrictions (an average of 20% of the constriction gesture is done by the head) and uses for now a crude approximation of the speech/gesture coordination.

4.2 From gestures to appearance

The text-to-cued speech synthesis system sketched above delivers trajectories of a few fleshpoints placed on the surface of the right hand and face. 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 is first mapped onto the existing face and hand parameter space. A further appearance model using video-realistic textures is then added (see figure 4).

5. CONCLUSIONS AND PERSPECTIVES

The immense benefits of FMCS in terms of giving access to the language structure and speech comprehension to deaf people should be grounded on a deep understanding of its implementation by actual speakers. Although precise qualitative guidelines have been specified by Orvin Cornett, the FMCS is a living language whose phonetic structure is constantly enriched by cuers. The observation of cuers in action is thus a perquisite for developing technologies that will assist deaf people in learning the FMCS. Low rate transmission of MCS by watermarking actual audiovisual transmission as put forward by the ARTUS consortium should also benefit from a better understanding of the kinematics of the different segments involved in the production of MCS. The database recorded, analyzed and characterized here is currently been exploited within a multimodal text-to-FMCS speech system that will supplement or replace on demand subtitling for TV broadcasting or home entertainment.

6. ACKNOWLEDGMENTS

Many thanks to Yasmine Badsi, our MCS speaker for having accepted the recording constraints. We thank Frédéric Vandenbergh for the processing of the raw motion capture data. We acknowledge also Virginie Attina, for providing her cued speech expertise when needed. This work has been financed by the RNRT ARTUS.

REFERENCES

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

[2] G. Bailly and P. Badin. Seeing tongue movements from outside. In *International Conference on Speech and Language Processing*, pages 1913–1916, Boulder, Colorado, 2002.

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

[4] L. E. Bernstein, M. E. Demorest, and P. E. Tucker. Speech perception without hearing. *Perception and Psychophysics*, 62:233–252, 2000.

[5] R. O Cornett. Cued speech. *American Annals of the Deaf*, 112:3–13, 1967.

[6] R. O Cornett. 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):375–384, 1988.

[7] O. Engwall and J. Beskow. Resynthesis of 3d tongue movements from facial data. In *EuroSpeech*, Geneva, 2003.

[8] G. Gibert, G. Bailly, D. Beautemps, Eliséi F., and R. Brun. Analysis and synthesis of the 3d movements of the head, face and hands of a speech cuer. *Journal of the Acoustical Society of America*, submitted for publication.

[9] A. J. Hunt and A. W. Black. Unit selection in a concatenative speech synthesis system using a large speech database. In *International Conference on Acoustics, Speech and Signal Processing*, pages 373–376, Atlanta, GA, 1996.

[10] J. Jiang, A. Alwan, L. Bernstein, P. Keating, and E. Auer. On the correlation between facial movements, tongue movements and speech acoustics. In *Proceedings of International Conference on Speech and Language Processing*, pages 42–45, Beijing, China, 2000.

[11] J. Leybaert. The role of cued speech in language processing by deaf children: an overview. In *Auditory-Visual Speech Processing*, pages 179–186, St Jorioz, France, 2003.

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

[13] G. Nicholls and D. Ling. Cued speech and the reception of spoken language. *Journal of Speech and Hearing Research*, 25:262–269, 1982.

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

[15] L. Revéret, G. Bailly, and P. Badin. 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*, pages 755–758, Beijing, China, 2000.

[16] R. Uchanski, L. Delhorne, A. Dix, L. Braida, C. Reed, and N. Durlach. Automatic speech recognition to aid the hearing impaired: Prospects for the automatic generation of cued speech. *Journal of Rehabilitation Research and Development*, 31:20–41, 1994.

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