EVALUATION OF A SPEECH CUER: FROM MOTION CAPTURE TO A CONCATENATIVE TEXT-TO-CUED SPEECH SYSTEM

Guillaume Gibert*, **Gérard Bailly***, **Frédéric Eliséi***, **Denis Beautemps***, **Rémi Brun**†

* 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

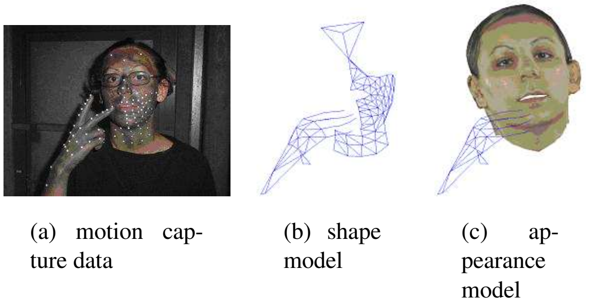
† Attitude Studio SA, 100 Avenue du Général Leclerc, 93692 Pantin France

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 ($R \sim 0.7$) (Yehia et al., 1998; Jiang et al., 2000) and this correlation is insufficient for recovering essential phonetic cues such as place of articulation (Bailly and Badin, 2002; Engwall and Beskow, 2003). 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 (Owens and Blazek, 1985) or in words or sentences (Bernstein et al., 2000). Manual Cued Speech (MCS) was designed to complement speechreading. Developed by Cornett (Cornett, 1967), 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). For more details on the French MCS (FMCS) system please consult http://retore.chez.tiscali.fr/LPC. Numerous studies have demonstrated the drastic increase of intelligibility provided by MCS compared to lipreading alone (Nicholls and Ling, 1982; Uchanski et al., 1994) and the effective facilitation of language learning using FMCS (Leybaert, 2003). 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.



(a) motion capture data

(b) shape model

(c) appearance model

Figure 1: From motion capture data to a videorealistic virtual speech cuer.

2. Motion capture data

We recorded the 3D positions of 113 markers glued on the hand and face of the subject (see figure1(a)) 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 (Badin et al., 2002).
- 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 face and hand

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 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 (Badin et al., 2002; Revéret et al., 2000) using the first principal component of different subsets of fleshpoints: we first subtract the contribution of the jaw rotation. The lips rounding/spreading gesture is estimated in the residual and then substract to the data. The proper vertical movements of upper and lower lips, of the lip corners as well as the movement of the throat are substracted in this order to the residual 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 frames (equal here to 2mm) is performed before modeling in order to whiten the training data.

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 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 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.

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 and *nmF* = *nmH* = 5 for the head and hand movements. 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 RMS modeling error for the position of the visible markers is equal to 2mm for the hand and 1mm for the face.

4. Further data analysis

Further data analysis was performed in order to verify if the cuer has effectively realized the recommended handshapes and hand positions with the consonants and vowels effectively produced. In the following all available data are considered. Movements and deformations of hand and face are regularized and reconstructed using the hand and face models described above. Globally the FMCS 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.

4.1. Recognizing hand shapes and consonants

We thus 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 fleshpoint 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 handshapes 2 versus 8.
- The absolute distance between the tip of the thumb and the palm in order to differentiate between handshapes 1 versus 6 and 2 versus 7.

4114 handshapes were identified and labeled. The 7 characteristic parameters associated with these target handshapes are then collected and simple Gaussian models are estimated for each handshape. The a posteriori probability for each frame to belong to each of the 9 handshape model can then be estimated. The recognition rate is quite high: we have a recognition rate of 98.78% . The ”errors” involve mainly confusions between the coding of mid-vowels (/e/ vs. /ε/) and omissions of the coding of glides in complex CCCV sequences.

4.2. Recognizing hand placements and vowels

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. Of the 4114 hand placements, 96.76% were identified with a total of 133 errors. There are mainly two identified sources of errors:

- the main source of error comes from the hand placement 1 (side). This hand placement is used for a consonant followed by another consonant or a schwa and undershoot of this short target occurs very often (i.e. the cuer only points to side but does not reaches it).
- hand placement 0 displays a large variance and hand placements 1 (side) and 4 (cheeks) realized too far away from the face are sometimes captured by the Gaussian model for hand placement 0.

We exhibit in figure 2 an example of the time course of these probability functions over the first utterance of the corpus together with the acoustic signal.

4.3. Phasing speech and gestures

In the present study, the data also gathers very precious data on phasing relations between speech and hand gestures. We analyzed the profile of hand shape and hand placement gestures (we verify manually the pre-segmentation done before using our MOTHER OPENGL©animation software (Revéret et al., 2000)) 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. A sketch of the profiles for CV sequences is presented in Figure 3: 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 and ”isolated” consonants, 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.

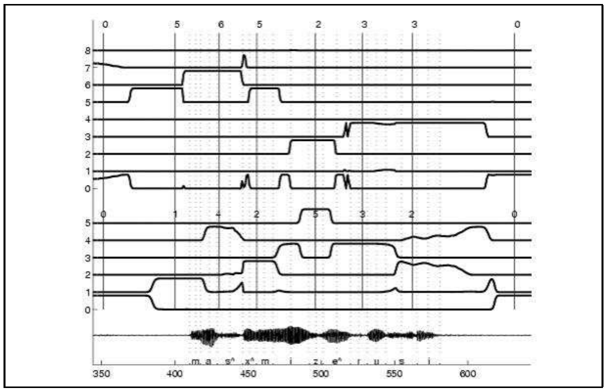


Figure 2: Recognition of the hand shapes (top) and hand placements (bottom) by simple Gaussian models. The vertical lines show hand targets together with the required hand shapes and placements.

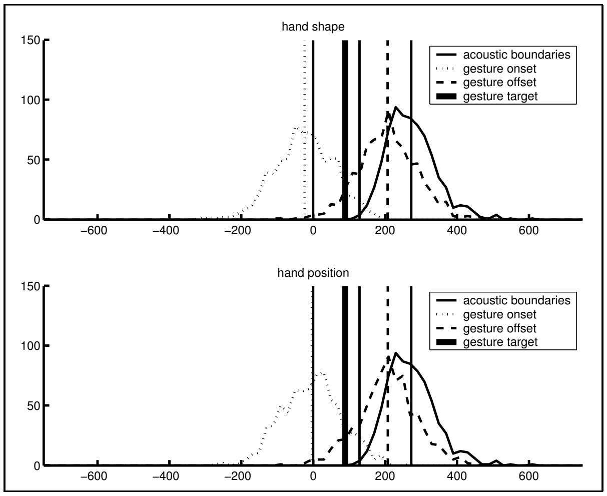


Figure 3: Phasing gestures with reference to the CV acoustic segments.

5. 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 (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 gestures 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. 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.

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. 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 (Bailly et al., 2002) is applied that consider the onset of each dikey as the intended target: a linear interpolation of hand placement applied gradually 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.

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 (see figure 1(b)). 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 1(c)).

6. Conclusions and perspectives

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.

7. 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.

8. References

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

Bailly, G. and P. Badin, 2002. Seeing tongue movements from outside. In *International Conference on Speech and Language Processing*. Boulder, Colorado.

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

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

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

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

Hunt, A. J. and A. W. Black, 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.

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

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

Minnis, S. and A. P. Breen, 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.

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

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

Revéret, L., G. Bailly, and P. Badin, 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.

Uchanski, R., L. Delhorne, A. Dix, L. Braida, C. Reed, and N. Durlach, 1994. 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.

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