Evaluation of a Speech Cuer: From Motion Capture to a Concatenative Text-to-cued Speech System

Guillaume Gibert[1,2], Frédéric Elisei[1] & Gérard Bailly[1]
[1] Gipsa-lab, DPC, 46 av. Félix Viallet, 38031 Grenoble cedex, France
[2] INSERM, U821, Lyon, F-69500, France
Guillaume.Gibert@inserm.fr; {Frederic.Elisei;
GerardBailly}@gipsa-lab.inpg.fr

We present here our effort for characterizing the 3D movements of the right hand and the face of a French female producing Cued Speech. Cued Speech was designed by Cornett (Cornett 1967) for complementing speech reading with additional phonetic information provided by hand gestures for deaf people. Experimental results show a drastic increase in intelligibility and speech learning of hearing impaired people. 8 hand shapes code subsets of consonants and 5 hand positions on the face code subsets of vowels. Consonants and vowels within each subset are chosen as being easily identified by only facial cues.

The observation of cuers in action is a prerequisite for developing technologies that will assist deaf people in learning and using French Cued Speech (FCS). To study FCS, we recorded the 3D positions of 113 markers glued on the hands and face of a subject (see Figure 1 (a)) using a VICON® motion capture system with 12 cameras at 120 Hz. 3 corpora are designed: (a) a corpus of hand shapes transitions produced in free space. This corpus is used for building a statistical model of hand movements; (b) a corpus of visemes uttered without cueing. This corpus is used for building a statistical model of facial movements; (c) a corpus of 238 sentences uttered with cueing FCS. This corpus provides an extensive coverage of multi-represented French diphones, especially designed for acoustic concatenative speech synthesis. This unique resource constitutes a dense and robust access to FCS movements.

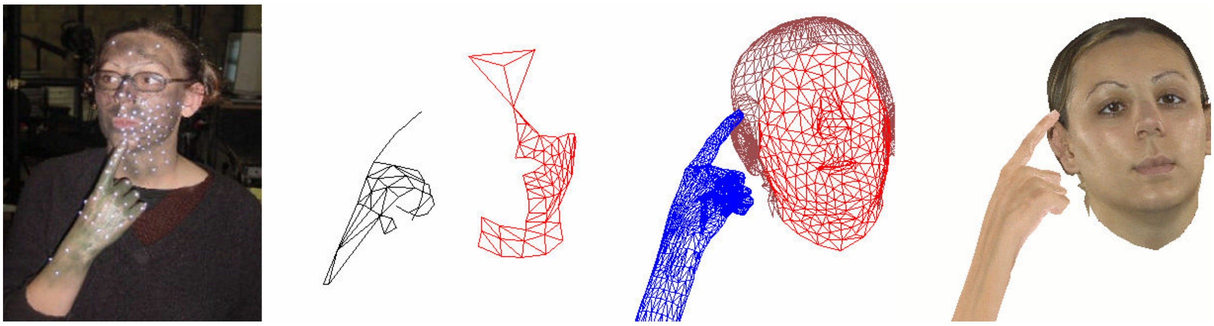

Figure 1 - from motion capture to a virtual cuer (left to right, (a) motion capture session, (b) motion capture data, (c) shape model, (d) appearance model).

We constructed linear and non-linear statistical models of hand and face movements with a precision close to the millimetre and with only 13 and 18 parameters for driving respectively the face and the hand (Gibert, Bailly *et al.* 2005). These models developed from corpora (a) and (b) are further used to clean up the FCS corpus (c) and give a continuous access to the hand and face gestures during cued speech production.

Further data analysis was performed to verify whether the cuer has effectively realized hand shape and hand position in accordance with the consonant and the vowels pronounced. We selected target frames in the vicinity of the relevant acoustic event and labelled them with appropriate key value. Characteristic parameters associated with these target hand shapes (distance between palm and finger tip for each finger as well as distances between finger tips) and positions (3D position of the longest finger in reference to the head) are then collected and simple Gaussian models are estimated respectively for each hand shape and position. The recognition rate is quite high: 98.78% for consonant keys and 96.76% for vowel keys. The mainly source of errors are coming from isolated consonant or vowels: isolated phones indeed force the coder to be faster and targets are undershot.

These Gaussian models are also used to study the temporal organization in the production of CV syllable for French Cued Speech. We analyzed the profile of hand shape and hand placement gestures in reference to the acoustic realization of speech segment they are related to. We concluded that both hand shape and hand position are realized well before the acoustic onset of the speech segment they are related to. Moreover, the hand shape and the hand placement gestures are highly synchronized.

Although the phonetic content of corpus (c) was designed initially for acoustic concatenative speech synthesis, we used our multimodal data for generating hand and face movements together with speech using concatenation of gestural and acoustics units. We consider 2 kinds of units: diphones for the generation of the acoustic signal and facial movements and “dikeys” for the generation of head motion as well as hand movements and articulations.

The text-to-cued speech synthesis system sketched above delivered trajectories of a few flesh points placed on the surface of the right hand and face (see Figure 1 (b)). High definition models of these organs are first mapped onto the existing face and hand parameter space (see Figure 1 (c)). A further appearance model using video realistic texture is then added (see Figure 1 (d)).

A first series of experiments have been conducted to evaluate the intelligibility of this virtual cuer with skilled deaf users of the French cued speech. The first evaluation campaign was dedicated to segmental intelligibility. We developed a specific test which mirrors the Modified Diagnostic Rime Test developed for French by Peckels and Rossi (Peckels and Rossi 1973). Mean intelligibility rate for “lipreading” condition is 52.36% whereas mean intelligibility rate for FCS condition is 94.26%. The results of the preliminaries perceptive tests show that significant linguistic information with minimal cognitive effort is transmitted by our system.

REFERENCES

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

Gibert, G., G. Bailly, *et al.*, 2005. Analysis and synthesis of the three-dimensional movements of the head, face and hand of a speaker using Cued Speech. *Journal of the Acoustical Society of America* **118**(2): 1144--1153.

Peckels, J. P. and M. Rossi, 1973. Le test de diagnostic par paires minimales. *Revue d'Acoustique* **27**: 245--262.