AVSP 2003 – International Conference
on Audio-Visual Speech Processing
St. Jorioz, France
September 4-7, 2003

Toward an audiovisual synthesizer for Cued Speech : Rules for CV French syllables

V. Attina, D. Beautemps, M.-A. Cathiard & M. Odisio

Institut de la Communication Parlée
CNRS UMR5009/INPG/Université Stendhal, Grenoble, France
attina@icp.inpg.fr

Abstract

Manual Cued Speech is an effective method used to enhance speech perception for hearing-impaired people. Thanks to this system, a speaker can clarify what has been said with the help of hand gestures. Seeing manual cues associated to lip shapes allows the cue receiver to identify speech elements unambiguously. A large amount of work has been devoted to Cued Speech effectiveness in visual identification, in the access to complete phonological representations and in language acquisition or reading and writing learning. No work aimed at investigating the temporal organization of Cued Speech production, i.e. the co-articulation of Cued Speech articulators. In this framework, the present paper presents an investigation of the temporal organization of hand cue presentation in relation to lip motion and the corresponding acoustic patterns in order to specify the nature of the syllabic structure of Cued Speech. Data reveal a clear advance of the hand on the sound and lip motion. Temporal coordination rules for French Cued Speech gestures are derived and an audiovisual synthesizer generating CV sequences in Cued Speech and based on these principles is presented.

1. Introduction

Speech communication is multimodal by nature : it is well-known that hearing people make use of visual information when the acoustic signal is difficult to perceive [1]. Many deaf people also depend on speechreading for communication. However visual information conveyed by speech articulators is not sufficient for a complete perception. Cued Speech (CS) was invented in 1967 to solve this problem by complementing speechreading with the hand [2] for American English language. Adapted to more than 50 languages [3], the French version so called "Langage Parlé Complété" or French Cued Speech appeared in 1980's.

In CS, while uttering, the speaker places one of her/his hand near the face to disambiguate lip shapes. Manual cues are formed along two parameters : hand placement and hand shapes. Placements of the hand code vowels whereas hand shapes code consonants. For French, CS uses 5 placements and 8 hand shapes for all the phonemes (Fig. 1). Phonemes with similar lip shapes are allocated among different cues. By contrast, phonemes easily discriminated on lips are associated to the same cue. In this way, a handshape at a specific placement around the face associated to a lip shape defines a single CV syllable. Special indications are defined for successive consonant groups and isolated vowels (Fig. 1).

Many studies have shown the efficiency of CS in improving speech perception for hearing-impaired people [4]. Moreover, early exposition to CS has been shown to be very efficient for language development of profoundly deaf children [5] (for a more detailed state of the art, see Leybaert's contribution to this volume).

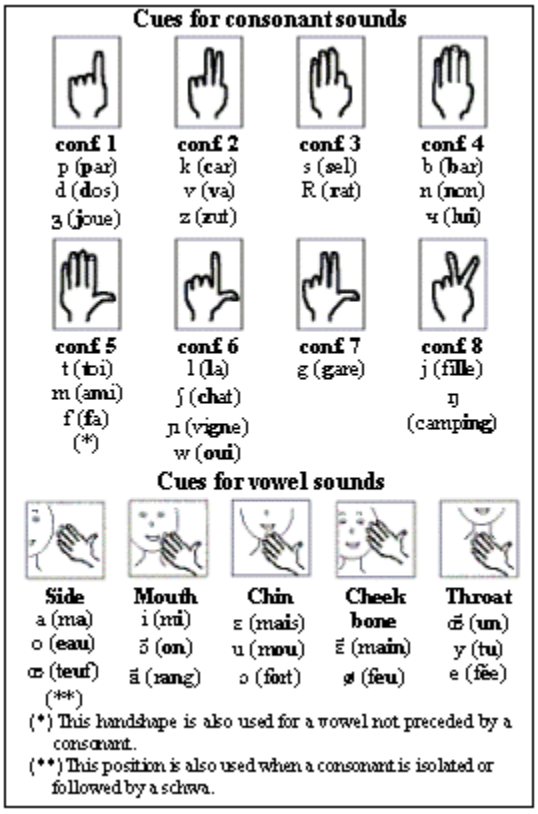

Figure 1: Cues for French language.

In perspective of Cued Speech synthesis, the analysis of CS production is crucial and in this framework the study of the coordinations between CS effectors (hand, fingers, face) in relation with the corresponding sound is a main issue. For this purpose, we investigated experimentally the temporal nature of the CS gestures relatively to orofacial movements from the analysis of a real CS speaker. A first experiment is devoted to the hand in comparison to lips and the corresponding acoustic events. The second experiment focused on the handshape formation relatively to the hand displacement. Finally an audiovisual synthesis system delivering CS CV sequences and based on the obtained rules for coordination of CS gestures is illustrated.

2. Determination of temporal rules for CS gestures

For both experiments, the same method has been applied. The approach is based on the analysis of biological signals through audio-visual recordings of a Cued Speech speaker pronouncing and coding nonsense syllabic French CV sequences (Consonant-Vowel).

Experiment 1 aims at exploring the hand gesture relatively to lip movements and sound. In this part there is no handshape change [6]. Experiment 2 focuses on the handshape formation with or without hand transitions [7].

Both experiments deliver a general pattern for French CS production for CV sequences. They provide rules for the temporal coordination of CS effectors used to improve the ICP first prototype of French CS synthesizer.

2.1. Method

2.1.1. Corpus

2.1.1.1 Experiment 1

The corpus was defined for hand gesture analysis. The handshape was fixed during the production of the whole sequences: there were no finger gesture during the production of each sequence. Only displacement of the hand between the 5 positions for CS were studied. The corpus was composed with [$CaCV_1CV_2CV_1$] sequences made of [m, p, t] consonants for C combined with [a, i, u, ø, e] vowels for V_1 and V_2 (the 1st syllable [Ca] was not analyzed). Finally, the whole corpus contains 20 sequences such as [mamamima], for each of the three consonants, so a total of 60 sequences. For each [$CaS_1S_2S_3$] sequence, the analysis focused on S_2 syllable (i.e. on transitions from the S_1 syllable toward S_2 and S_2 toward S_3).

2.1.1.2 Experiment 2

The corpus was elaborated to analyze finger gestures during handshape formation. The consonants were chosen to solicit only one finger in the change from a handshape to another one. This choice was motivated to facilitate data reading. Handshape formation was analyzed with two kind of sequences :

- Sequences implying only handshape modification at fixed placement for the hand. [mVC_1VC_2V] sequences with the same vowel (V = [a] or [ ]) were designed for consonant variation. The C_1 and C_2 consonants were [p] and [k], [s] and [b] or [b] and [m]. 10 repetitions of each sequence were recorded. The analysis focused on S_2 and S_3 syllables, resulting in 120 syllables (10 repetitions x 3 consonant groups x 2 vowels x 2 syllables)

- Sequences involving handshape modification simultaneously with hand transitions between target positions. [$mV_1C_1V_2C_2V_1$] sequences varied both vowel and consonant. The C_1 and C_2 consonants were [p] and [k], [ ] and [g], [s] and [b], or [b] and [m]. The V_1 and V_2 vowels were [a] and [u], [a] and [e] or [u] and [e]. 5 repetitions of each sequence were recorded. The analysis focused on S_2 and S_3 syllables, resulting in 120 syllables (5 repetitions x 4 consonant groups x 3 vowel groups x 2 syllables).

2.1.2. French Cued Speech speaker

The French Cued Speech speaker is a 36 year-old French female. She has been recommended by a speech therapist for the good hand visibility and fluidity during coding and for the good lip view in reception. She has been using Cued Speech at home with her hearing impaired child for 8 years. She graduated in French Cued Speech in 1996 and regularly translates into CS code at school.

2.1.3. Audiovisual recording

For both experiments, the recording was realized in a sound-proof booth, at 50 frames/second. A first camera in large focus was used for the hand and the face. A second one in zoom mode dedicated to the lips was synchronized with the first one. The lips were made-up in blue and colored marks were placed on the hand to follow the movement (Fig. 2). Finally, a blue mark was placed on the opaque glasses worn by the speaker as a reference point to the different measurements. The head was maintained fixed with a helmet to avoid mobility.

Figure 2: Image of the FCS speaker with axes in superimposition used for landmarks localization.

For Experiment 2, the same experimental setup was used in addition with a data collector glove to follow finger movements during the handshape formation. A colored mark was placed on the back of the glove to follow the displacement of the hand (Fig. 3).

Figure 3: Image of the CS speaker wearing the data-collector glove with superimposition of colored marks and axis used for the analysis

2.1.4. Data processing

For both experiments, the automatic extraction system by image processing developed at ICP [8] provided a set of lip parameters every 20 ms. We chose to explore the temporal evolution of the between-lips area, which is a

pertinent parameter to characterize labial forms under face angle view. In synchrony with lip area parameter, the audio signal was digitalized. We extracted the x and y coordinates of the hand mark placed near the wrist for Experiment 1 and of the mark placed on the glove for Experiment 2 (at a 50 Hz frequency). In addition, for Experiment 2, the data glove provided raw data values for all the 18 sensors.

Finally, the process provided synchronous signals : the acoustic signal, lip area values, x and y trajectories of the colored mark of the hand (Fig. 4) or of the data glove and in case of Experiment 2, angle values from the glove for the representative sensor (the one measuring the movement of the finger studied) (Fig. 5).

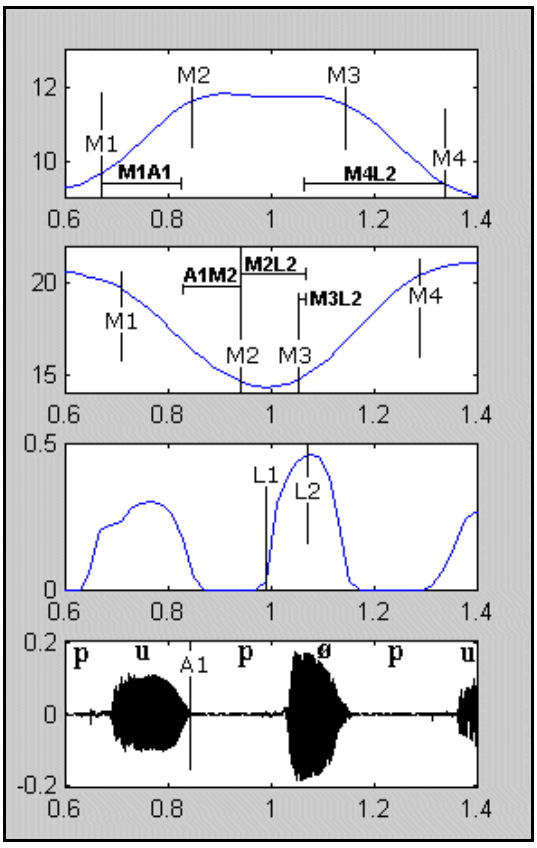

Figure 4: From top to bottom: (1). x (cm) and (2) y (cm) trajectories of the hand for [pupøpu] part of a [papupøpu] sequence (50 Hz). (3). Time course of lip area S (cm^2) (50 Hz); (4). The corresponding acoustic signal (22050 Hz). On each signal, labels and temporal intervals used for the analysis.

For the analysis of each [$S_1S_2S_3$] syllabic sequence (Fig.4 and Fig. 5), the acoustic signal was labeled at the consonant onsets of S_2 and S_3 (A1 and A2 labels). The lip gesture, the hand transition and the finger movement were manually labeled at the acceleration peaks [9]. For lip area, L1 marks the beginning of the S_2 vocalic gesture and L2 is for the reached target. M1 is the beginning of the hand gesture toward the position corresponding to S_2, M2 the reached target position (coding S_2) maintained until M3, time where the hand begins the gesture toward the following position for S_3 coding, M4 corresponding to the S_3 reached target. Finally, for Experiment 2, D1 marks the beginning of the finger gesture corresponding to the consonant of S_2 syllable and D2 the end of the gesture for handshape formation. D3 marks the beginning of the finger gesture for S_3 syllable and D4 the end of the handshape formation for S_3.

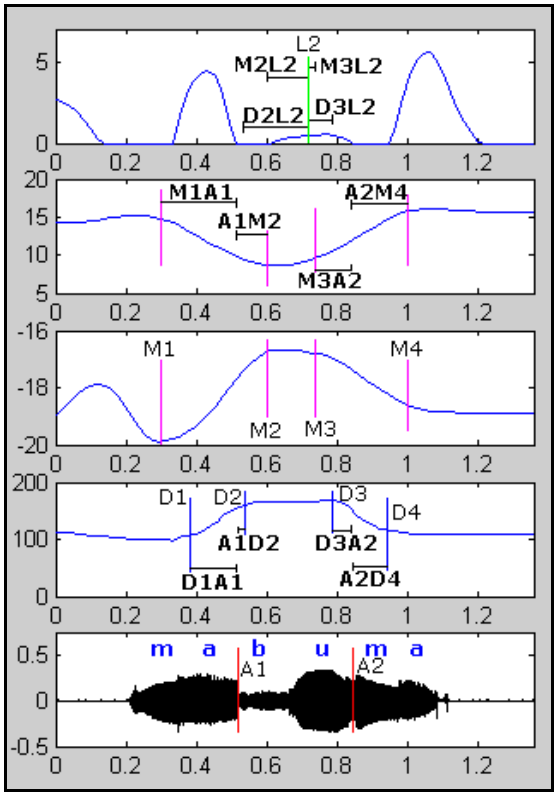

Figure 5: *From top to bottom for [mabuma] sequence : (1). Time course of lip area S (cm^2) (50 Hz) ; (2). x (cm) and y (cm) trajectories of the hand (50 Hz) ; (3). Temporal trajectory of the raw-data from the thumb first articulation glove sensor (64 Hz); (4). The corresponding acoustic signal (22050 Hz). On each signal, labels and temporal intervals used for the analysis.*

2.2. Results

Different temporal intervals derived from the events labeled on each signal were used to analyze the coordination between the FCS effectors (see these intervals on Fig. 4 and Fig. 5).

2.2.1. Experiment 1

Results showed a clear advance of the hand gesture on both the lip gesture and sound emission. The hand began its movement toward a target position 239 ms before the consonant acoustic onset (M1A1 interval) and reached the position 37 ms after (A1M2). Thus the hand target is attained during the first part of the consonant, quasi synchronously with it and largely before the lips formed the vocalic target (256 ms average value for M2L2). The hand maintained its position during the whole consonant and began to move toward the next position 51 ms before the vocalic lip target realization (M3L2).

2.2.2. Experiment 2

Results showed an advance of hand and finger movements on lips motion and sound events.

For sequences with handshape modification only, the finger began to move clearly before the acoustic onset of the consonant (mean value of 123.8 ms for D1A1) and finished the handshape formation at the beginning of the acoustic consonant (mean value of 46.5 ms for A1D2).

For sequences with both handshape change and hand transitions, the handshape began its formation before the acoustic consonant silent instant (mean value of 171.5 ms for D1A1) and was completely formed at the beginning of the syllable (A1D2), during the first part of the consonant. Concerning the hand, it began its movement before the finger gesture, so largely before the acoustic onset of the consonant (mean value of 205.1 ms for M1A1) and reached its position just after the end of the finger gesture during the consonant (mean value of 33.2 ms for A1M2) and so largely before the vocalic lip target (mean value of 172.2 ms for M2L2). Then, the hand left its position toward a new target position before the vocalic lip target (preceding vowel) (mean value of 43.4 ms for M3L2) and reached the position during the consonant (A2M4).

2.2.3. Summary of the 2 experiments

We obtained a noticeable coherence with results of the 2 experiments (see Fig. 6 for a general pattern). To sum up, concerning hand position, it was observed:

- The displacement of the hand toward its position begins before the consonant acoustic onset of the CV syllable. This implies that the gesture begins in fact during the preceding syllable, i.e. during the preceding vowel;
- The hand target is attained at the beginning of the acoustic consonant onset.

These results reveal the *anticipatory* gesture of the hand motion over the lips since the hand placement gesture covers the whole syllable duration, with a temporal advance over the vocalic speech gesture.

Finally, it was observed from the data-collector glove that the handshape is completely formed at the instant where the hand target position is reached. Moreover we noticed that the handshape formation gesture uses a large part of the hand transition duration.

3. Discussion: The Cued Speech co-production

We now explain how we consider the two Cued Speech components within the framework of speech control for the elaboration of a quantitative control model for Cued Speech production.

For transmitting the consonant information the control type is a figural one, i.e. the postural control of the hand configuration (fingers configuration). The type of control for transmitting the vowel information is a goal-directed movement performed by the wrist carried by the arm. These two controls are linked by an in-phase locking.

On the other hand, for speech, there are three types of control [10] [11].

- The mandibular oscillation control is the control of a cycle, self-initiated and self-paced. This control transmits the information of the syllabic rhythm, which is the basic carrier of speech.

The carried articulators (tongue and lower lip) together with their coordinated partners (upper lip, velum and larynx) achieve two types of transmitted information:

- The vowel information is produced by a global control of the whole vocal tract, i.e. a figural or postural motor control type.
- The consonant information is produced by the control of contact and pressure performed on local parts along the vocal tract.

The mandibular and vowel controls are coupled by an in-phase locking. The consonantal control is typically in phase with the vowel for the initial consonant of CV syllable. But it can be out-of-phase for the coda consonant in CVC syllable. And finally consonants gestures in clusters within the onset or the coda can be in phase (e.g. [psa] or [aps]) or out of phase ([spa] or [asp]).

As for speech, Cued Speech vowel and consonant depend on the wrist-arm *carrier* gesture, which is analogous to the mandibular rhythm. The control of the vowel *carried* gesture is a goal-directed movement which aims at a local placement of the hand on the side of the face. Whereas the consonant *carried* gesture is a postural (figural) one. Thus the two types of control in CS are inversely distributed in comparison to speech: the configuration global control of the speech vowel corresponds to a local control in CS whereas the local control for the speech consonant corresponds to a global control in CS.

Thus once speech rhythm has been converted into CS rhythm (that is a general CV syllabification with some cluster specificities), the two carriers (mandibula and wrist) can be examined with respect to their temporal coordination, i.e. phasing. This CV re-syllabification means that every consonantal CS gesture will be in-phase with the vocalic one, which is not always the case in speech for languages with CVC or out-of-phase consonant clusters. Contrary to speech, the CS consonant gesture never hides the beginning of the in-phase vocalic gesture (cf. Öhman's model [12]). Concerning the phasing of the two carried vowel gestures, our experiments made clear that the CS vowel gesture did anticipate the speech vowel gesture.

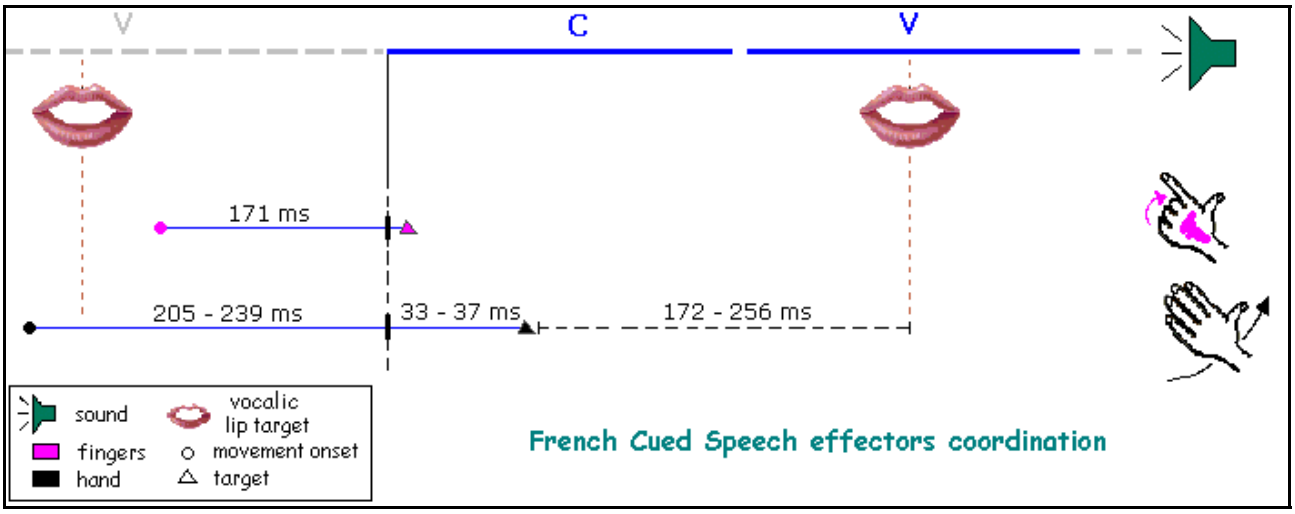

Fig. 6: General scheme of coordination for hand, fingers and lips in relation with sound for French Cued Speech production

4. The ICP Cued Speech audiovisual synthesizer

A Cued Speech synthesizer is designed to translate a chain of phonemes marked in time into Cued Speech. The resulting cues are added to a speaker face on a visual output. In the autocuer system developed by Cornett et al. [3], the cues were obtained from the acoustic speech recognition process of the pronounced word and displayed on group of LEDs on glasses worn by the speechreader. The whole process involves a delay of 150 to 200 ms for the cue display in comparison to the production time of the corresponding sound. This system designed for isolated words attained 82% of correct identification.

In the system of automatic generation of Cued Speech developed by Duchnowski et al. [13] for American English language, pre-recorded handshape photos are superimposed to the video recorded speaker face. The cues and the instant of presentation were derived from speech recognition process of the associated video acoustic signal. The whole process involves a delay of 2 s for the display of the video speaker face augmented with the Cued Speech hand modality. Scores of correct word identification reached the mean value of 66% and were higher than the 35% obtained with speechreading alone; but were still under the 90% level obtained with manual Cued Speech. This 66% mean score was obtained for the more efficient display, called "synchronous", in which 100 ms were allocated to the hand target position and 150 ms to the transition between two positions. Moreover, the instant of cue display was advanced by 100 ms relatively to the start time determined by the acoustic speech recognizer. This advance was fixed empirically by the authors.

At the ICP, an audiovisual synthesizer delivering CV sequences in Cued Speech was developed. The Cued Speech modality was integrated in the ICP audiovisual synthesizer made of a virtual talking head system. The ICP talking heads consist of a geometric 3D modelling of a speaker face by a set of points (mesh) on which a realistic texture is applied. A statistic factor analysis of the real face movements is used to extract a reduce set of parameters that control linearly the model i.e. the control parameters used for facial animation [14]. Control parameters are not linearly correlated and have an articulatory explanation: for example two of them are principally related to the jaw and three of them to lips movement. For the Cued Speech modality, pre-recorded photos of Cued Speech handshapes are superimposed on the virtual talking head (Fig. 7).

Figure 7 : View of the ICP virtual talking head with a CS handshape in superimposition

In the production of a CV sequence where the acoustic beginning of the consonant and of the vowel are temporally marked (from the output of a prosodic model [15]), a module generates the temporal evolution of the virtual talking head control parameters from rules based on targets for each of the phonemes and from a

coarticulation model that simulates speech context variability [16]. By the way another module controls the movement of a reference point of the back of the hand common to all the handshape photos. For this reference point, (x_c, y_c) targets are fixed for each of the five CS hand positions and the x and y trajectories between two (x_c, y_c) targets are derived from a sinusoidal modelling. The period of the sinus equals to 2 times the duration of the transition so that the derivate is null at the targets. The instant of the CS target positions are obtained from rules derived from the previous studies: The hand reaches the vowel (x_c, y_c) target position at the instant of the acoustic beginning of the consonant and maintains it until the acoustic beginning of the vowel is attained at which instant the hand starts its movement toward the target position of the next CV syllable. When necessary, the handshape change occurs at the middle instant of the hand transition.

Finally, a module based on diphone concatenation following a TDPSOLA technique [17] generates the associated synthesized acoustic signal from the CV chain temporally marked.

The evaluation of the whole system, based on the analysis of the identification of synthesized non sense CV sequences, is still under process and will be presented at the conference.

5. Acknowledgements

Many thanks to Martine Marthouret, speech therapist at Grenoble hospital, for the helpful discussions; to Mrs G. Brunnel, the Cued Speech speaker for having accepted the recording constraints, to C. Savariaux and A. Arnal for the recording technical support and to B. Celle for the first version of the demonstrator. This work is supported by the Remediation action of the French Research Ministry "programme Cognitique", a "Jeune équipe" project of the CNRS (French National Research Center) and a BDI grant from CNRS.

6. References

[1] Reisberg, D., McLean, J. and Goldfield, A. "Easy to hear but hard to understand: a lipreading advantage with intact auditory stimuli". *Hearing by Eye: The Psychology of LipReading*. B. Dodd and R. Campbell. Hillsdale, New Jersey, Lawrence Erlbaum Associates**:** 97-113, 1987.

[2] Cornett, R. O.. "Cued Speech." *American Annals of the Deaf, Vol. 112, 1967, p. 3-13.*

[3] Cornett, R. O. "*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.

[4] Nicholls, G. and Ling, D. "Cued Speech and the reception of spoken language." *Journal of Speech and Hearing Research*, Vol. 25, 1982, p 262-269.

[5] Leybaert, J. "Phonology acquired through the eyes and spelling in deaf children." *Journal of Experimental Child Psychology*, Vol. 75, 2000, p 291-318.

[6] Attina V., Beautemps D. & Cathiard M.-A., "Coordination of hand and orofacial movements for CV sequences in french Cued Speech", *Proc. of ICSLP*, Denver, 1945-1948, 2002.

[7] Attina V., Beautemps D. & Cathiard M.-A., "Temporal motor organization of Cued Speech gestures in the French language", *Proc. of 15th ICPHS, Barcelona*, 2003.

[8] Lallouache M. T., "Un poste visage-Parole couleur. Acquisition et traitement automatique des contours des lèvres", *PhDThesis, INP Grenoble*, 1991.

[9] Schmidt, R. A., "*Motor Control and Learning: A Behavioral Emphasis"*. Champaign, IL, Human Kinetics Publishers, 1988.

[10] Vilain A., Abry C. & Badin P.. "*Coproduction strategies in French VCVs: Confronting Öhman's model with adult and developmental articulatory data."* 5th Seminar on Speech Production. Models and Data, 1-4 May 2000, Kloster Seon, Bavaria, 81-84, 2000.

[11] Abry C., Cathiard M.-A., Vilain A., Laboissière R. & Schwartz J.-L. "Some insights in bimodal perception given for free by the natural time course of speech production." In G. Bailly, P. Perrier & E. Vatikiotis-Bateson (Eds.), *Festschrift Christian Benoît*, MIT Press. (to appear)] :

[12] Öhman, S. E. G. "Numerical model of coarticulation." *Journal of the Acoustical Society of America* **41**, 1967, p. 310-320.

[13] Duchnowski, P., D. S. Lum, J. C. Krause, M. G. Sexton, M. S. Bratakos and L. D. Braida. "Development of speechreading supplements based on automatic speech recognition", *IEEE Transactions on Biomedical Engineering*, 47(4), 487-496, *2000*.

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

[15] Morlec, Y., Bailly, G., and Aubergé, V. "Generating prosodic attitudes in French: data, model and evaluation". *Speech Communication, 33(4), 2001, p. 357-371*.

[16] Elisei F., Odisio M., Bailly G. & Badin P., "Creating and controlling video-realistic talking heads". *Proc. of AVSP*, 90-97, 2001.

[17] Bailly, G. and Alissali, M. "COMPOST: a server for multilingual text-to-speech system". *Traitement du Signal, 9(4): 1992, p. 359-366*.