A PERCEPTUAL DESYNCHRONIZATION STUDY OF MANUAL AND FACIAL INFORMATION IN FRENCH CUED SPEECH

Emilie Troille[1,2], Marie-Agnès Cathiard[2] & Christian Abry[2]

[1]GIPSA-Lab-ICP, UMR 5216 CNRS-INPG-Université Stendhal
[2]Université Stendhal - BP25 - 38040 GRENOBLE Cedex 9 FRANCE
troille@icp.inpg.fr, marieagnes.cathiard@u-grenoble3.fr, christian.abry@u-grenoble3.fr

ABSTRACT

French Cued Speech, adapted from American Cued Speech, disambiguates lipreading by a manual code of *keys* allowing the deaf to recover a more accurate phoneme identification. Using movement tracking of manual and facial actions coproduced in Cued Speech (CS), Attina et al. [4] evidenced a significant *anticipation of the hand over the lips*. In the present study we tested the natural temporal integration of this bimodal hand-face communication system, using a desynchronization paradigm in order to evaluate the robustness of CS to temporal decoherence. Our results obtained with 17 deaf subjects demonstrate that hand gestures can be delayed relative to the lips without consequences for perception, as long as this delay does not push the hand outside the visible articulatory phase of the consonant constriction state. Perceptual coherence or recomposition of coherence (recoherence) depends crucially on the compatibility of hand and mouth states, i.e. on the timing patterns evidenced in our preceding production studies.

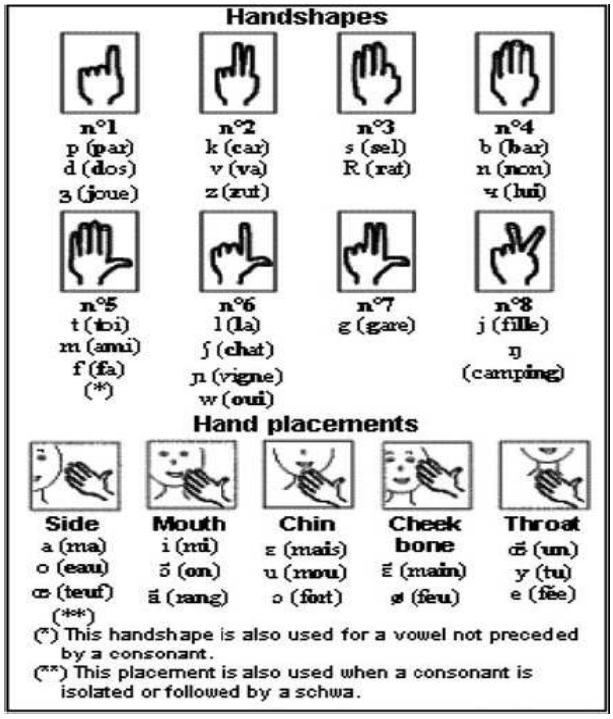
Keywords: Deafness, Cued Speech Production and Perception, Multimodality, Desynchronization

1. CUED SPEECH PRODUCTION

Cued Speech, created by Cornett in 1967 [1], was adapted to French in 1977, dubbed *Langue française Parlée Complétée*. As original Cued Speech, French CS disambiguates lipreading with the help of an augment, a manual code of *keys* coproduced around the face, in order to allow the deaf to recover the phonological segment message instead of ambiguous visemes.

CS supplements lipreading by two types of manual cues, all presenting the dorsal part of the hand: (i) the shape of the hand, displaying different finger configurations, informs about the consonants; (ii) while the hand position around the face codes the vowels (figure 1).

Figure 1: Hand configurations and positions in French Cued Speech (from [4])



The rationale is to put under the same hand shape or hand position a cluster of consonants or vowels which differ maximally at their mouth configuration (e.g. n°5: [t, m, f]). Hence sounds which are visually similar will be coded differently by the hand: e.g. the labial viseme for [p, b, m] will be respectively coded by keys n°1, 4 and 5. The same for the use of the various hand positions for coding vowels. It must be emphasized that this method allows full oral language phoneme perception only by the *combination* of manual keys together with lip forms. Contrary to fingerspelling, which is a completely different segmental (orthographic) code, it is proven that CS-Cuers can keep close to the natural speech rhythm, even in languages with consonant clusters (like English or French), once they are resyllabified into CV trains. Summarizing the organization of a CV syllable in CS: it associates a hand-shaping for the consonant

while the hand is targeting toward the position around the face for the vowel, the whole making the key for a specific CV. E.g.: syllable [de] (*dé* "thimble" in French) combines the consonant configuration n°1 with the position at the chin for the vowel: hence it will be produced by a pointing movement with the index toward the chin.

Beyond these CS principles, how are these code combinations actually coproduced and controlled? And how is the resulting flow of bimodal mouth-hand perceptual information integrated by the deaf? According to Cornett, CS is simply «[…] a time-locked system; that is, the cues must synchronize with the spoken sounds. Every cue is essentially a hand movement that is timed relative to the sound» [2]. Attina and al. [3, 4] were the first to achieve a series of studies with a movement and shape tracking system for manual and facial actions during CS coproduction. They disclosed for all their four coders a clear *anticipation* of hand movements on the lips (up to 200 ms), with the formation of the hand shape and its target position being completed quite synchronously with the beginning of the acoustic constriction state of the consonant, actually a *phase-locked* rather than a «time-locked» CS coordination. So CS appears to be the reverse of a system where lip information is supplemented by the hand, as designed by Cornett. On the contrary, the evidence given by these CS production studies is now that «the hand placement first gives a set of possibilities for the vowel, the lips then delivering the uniqueness of the solution» [3]. In our words: «hand first, lips next».

The temporal organization of hand and mouth gestures in CS production being outlined, the issue we address now tests specifically how this time course is taken into account during its perception by deaf people. In this direction, Attina et al. [5] undertook a first perceptual experiment which consisted in a gradual delivery of the information flow of a CV syllable (inserted in a carrier sentence) using a *gating* paradigm. Their study indicated that the natural anticipation of the hand over the lips, as found in the analysis of coder productions, was exploited in a gradual way along its time course by deaf subjects.

2. DESYNCHRONIZATION EXPERIMENT

The present experiment was carried out to study perceptual integration of the manual and lip information in French CS using a *desynchronization* paradigm. Studies of audio-visual desynchronization have shown that in speech, a change in lead is more tolerated than a change in lag [6, 7, 8]. Knowing that in CS the hand leads the face, a video of a French professional coder was recorded, and a progressive lag in her hand was edited. More specifically, the aim of this study was to test if the occurrence of the completion of the hand shape at the beginning of the constriction state of the consonant could be delayed, and with what consequences. Could this event be delayed outside the constriction phase, far beyond its end? Or could this regular CS *phase-locking* be maintained to some extent, being just adaptively flexible within the constriction phase?

2.1. Material

[paCaba] sequences were used, where C was chosen among [b, d, n, p] consonants. The C hand configuration was n°1 for [p, d], and n°4 for [n, b], all produced on the side position for [a]. So four sequences were obtained: [pababa], [padaba], [panaba], [papaba]. The choice of a side position, far enough from the face, would facilitate image editing for desynchronization (by cutting easily the image vertically in two parts: see below). As regards the consonant choice of [p, b], they visually clearly contrast from [d, n].

2.2. Recording and signal analysis

The French cuer was audiovisually recorded in an anechoic room, with 2 cameras, one for a close-up of her lips (for precise detection of lip area with the *ICP Lip-Shape Tracker System* [9]), and the other camera for a wider shot of the face and the hand. We processed our signals in order to check that the pattern of hand anticipation of our cuer was quite in conformity with the general production pattern evidenced by Attina et al. [4]. On figure 2 are shown, presented in synchrony with the lip area function, the results of our image-to-image visual inspection of the stability (no finger move) for the two hand shapes. Their transition phase is simply indicated by a linear interpolation between the two steady states. It is visible that the formation of the consonant key begins during the second half of the preceding vowel (i.e. the first [a] in our sequences) and that it is completed about the beginning of the constriction phase of the consonant (40 ms after for [pababa], and about its onset for [panaba]). For [pabapa], the labial closure of [b] (zero lip area) lasts 160 ms (images 52 to 56); and for [panaba],

the duration of [d], measured on its acoustic spectrogram, is 200 ms (corresponding to images 62-67).

2.3. Visual test

As control sequences, [papaba] and [padaba] were not desynchronized. For [pababa] and [panaba], we divided each image vertically in two zones, in order to separate the face from the hand. Then we delayed by 40 ms steps the image of the hand relative to the face. 10 steps were tested up to +360 ms. For a 360 ms lag, the finger configuration of [p] in [panaba] falls in synchrony with another labial form, that of [n]. We hypothesized that this would possibly generate a [d] percept, this consonant having the same finger configuration as [p] and the same lip shape as [n]. In the same way, the [pababa] sequence, with the 360 ms lag, would be identified [papaba], lip shapes being the same. After editing these sequences with *Adobe Premiere Pro*, the perceptual test was designed under *Multimedia Toolbook*, with 10 desynchronized sequences for [panaba], and 10 for [pababa], plus the two control sequences [papaba] and [padaba]. All these sequences were repeated 5 times and presented in random order. We wondered when the two possible switches in categorization, [n]/[d] and [b]/[p], would occur along the 10 sequences after an increase in desynchronization.

18 severely or profoundly deaf subjects practicing French CS, 6 women and 12 men, from 12 to 34 years of age (mean: 22 yrs, 1 month) were volunteers. The test was visual only. Their task was to identify the second syllable among 4 choices: “ba”, “da”, “na” or “pa”. Subjects read test instructions, and started with a familiarization phase containing the 4 synchronized sequences. Then the test sequences were displayed and they gave for each an immediate response (what second syllable did you perceive?) by clicking on one of the four choice boxes.

2.4. Results

Identification scores pooled for 17 subjects are presented on figure 3 for each desynchronization step. (The results of the 18th subject were not plotted since, for whatever reason, he had a uniform “pa” response, in all conditions.) In the four presentations of synchronized sequences ([papaba], [padaba] controls; and [pababa], [panaba]), the second syllable was correctly identified (98% on average). Whereas responses to desynchronization will clearly focus on 2 choices over 4 (figure 3). So what about these desynchronized sequences?

In [pababa] (figure 3), [ba] identification remained high (above 88%), up to a +120 ms delay of the hand. A step farther (+160 ms), the consonant finger key had glided outside the consonant constriction phase, and the identification switched to “pa” crossing over at 45% with (Probit fitting for “pa” gives the 50% boundary at +184 ms). Beyond this boundary, a new stable coherence finally emerged, i.e. [papaba] when approaching the lip area climax of the second [a] vowel (above 90%).

In the same way, in [panaba], correct identification [na] was maintained (above 92%) as long as the hand delay did not exceed +200 ms, which covered the consonant constriction phase. Beyond, identification decreased quite slowly, in this case. But it never reached decisively a new percept, with a maximum “da” score at about 60% (like for the “na” curve, logistic fitting is not appropriate). In this respect the mouth shape for [na] is obviously less informative for the constriction phase of the consonant than closed lips for [pa] or [ba], since [da] constriction does not occur at the lips, but behind the teeth.

3. COHERENCE, DECOHERENCE, RECOHERENCE

As concerns the stability of the two test sequences [pababa] and [panaba] along a quite important first range of desynchronization, such a stability can be accounted if the phase-locking of the consonant key to the consonant constriction state − what we observed in CS natural production − is loose enough for tolerating adaptation. This range of tolerance does work as long as the consonant finger key does not cross the coherence boundary, i.e. the end of the consonant constriction state. For the two overall patterns, when desynchronization is pushed forward, [panaba] shows a coherence ending in a decoherent state, while [pababa] meets a coherence-decoherence-recoherence story.

In this study we tested the natural temporal integration of a bimodal communication system with the face and the hand, French Cued Speech. Using a desynchronization paradigm we were able to evaluate the robustness of CS to temporal decoherence. Perceptual coherence or recomposition of coherence (recoherence) depends crucially on the compatibility of hand and mouth

states, i.e. on the timing patterns evidenced in our preceding production studies. In conclusion, the present study supports the claim that the natural anticipation of a manual CS gesture, coding the consonant to come, can be delayed without consequences for perception, as long as this delay does not unhinge the natural articulatory timing of the consonant.

Acknowledgments: To the cuer and deaf subjects and to C. Savariaux for technical assistance. This work was supported by a grant from Région Rhônes Alpes (Cluster 11).

4. REFERENCES

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

[2] Cornett, R.O. 1994. Adapting Cued Speech to additional languages. *Cued Speech Journal*, V, 19-29.

[3] Attina, V., Beautemps, D., Cathiard, M.-A., Odisio, M. 2004. A pilot study of temporal organization of Cued Speech in production of French syllables: Rules for a Cued Speech synthesizer. *Speech Communication*, 44, 197-214.

[4] Attina, V., Cathiard, M.-A., Beautemps, D. 2006a. Temporal measures of hand and speech coordination during French Cued Speech production. In S. Gibet, N. Courty, J.-F. Kamp (Eds.): *GW 2005*, Lecture Notes in Artificial Intelligence, Springer-Verlag, 3881, 13-24.

[5] Attina, V., Cathiard, M.-A., Beautemps, D. 2006b. French Cued Speech: from production to perception and vice-versa. *Handicap 2006*, 7-9 June, Paris.

[6] Dixon, N. F., Spitz, L. 1980. The detection of audiovisual desynchrony. *Perception*, 9, 719-721.

[7] McGrath, M., Summerfield, Q. A. 1985. Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. *J. Acoust. Soc. Am.*, 77(2), 678-685.

[8] Cathiard, M.-A., Tiberghien, G. 1994. Le visage de la parole: une cohérence bimodale temporelle ou configurationnelle. *Psychologie française*, Special issue "La reconnaissance des visages", 39(4), 357-374.

[9] Lallouache M. T. (1991). Un poste visage-Parole couleur. Acquisition et traitement automatique des contours des lèvres. PhD, INP Grenoble.

Figure 2: Time course of lip area (squares) and hand shape steady states interpolated (diamonds) in [pababa] (left) and [panaba] (right) sequences (40 ms between images). The diamond line interpolation corresponds to the synchronized hand, while the dotted line indicates a desynchronization of +360 ms.

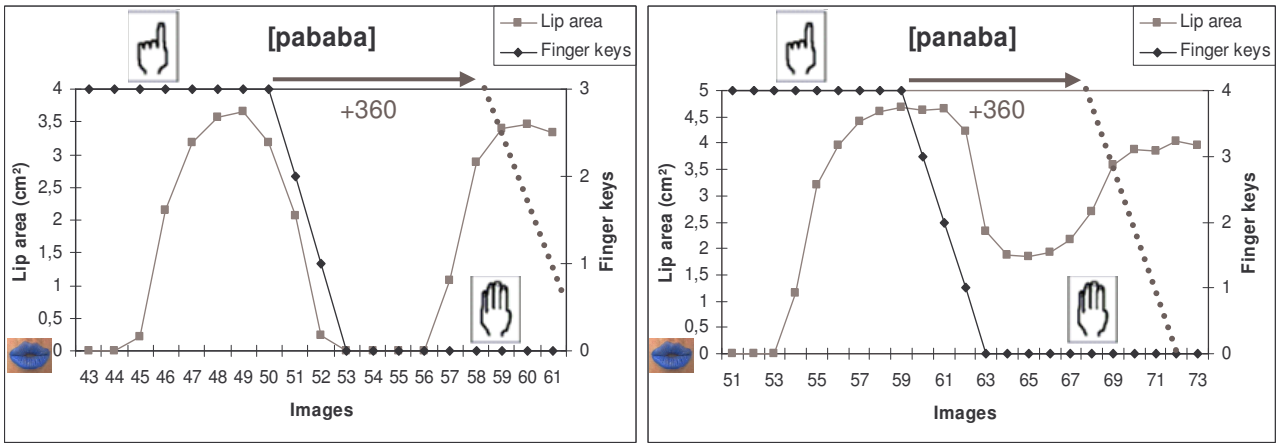


Figure 3: Identification scores for the second syllable of [pababa] (left) and [panaba] (right) along a 10 step-desynchronization continuum.