Temporal Organization of Cued Speech Production

Denis Beautemps, Marie-Agnes Cathiard, Virginie Attina, Christophe Savariaux

CONTENT

List of figures

List of tables

Erreur ! Aucune entrée de table d'illustration n'a été trouvée.

Index

Chapter 1. Temporal Organization of Cued Speech Production

D. Beautemps, M.-A. Cathiard, V. Attina, C. Savariaux, and A. Arnal

1.1 Introduction

Speech communication is multi-modal by nature. It is well known that hearing people use both auditory and visual information for speech perception (Reisberg, McLean et al. 1987)[1]. For deaf people visual speech constitutes the main speech modality. Listeners with hearing loss who have been orally educated typically rely on speech-reading based on lip and facial visual information. However due to the similarity in the visual lip shapes of speech units lip-reading alone is not sufficient. Even the best speech-readers do not identify more than 50 percent of phonemes in nonsense syllables (Owens and Blazek 1985) or in words or sentences (Bernstein, Demorest et al. 2000).

This chapter deals with Cued Speech, a manual augmentation for lipreading visual information. Our interest in this method was motivated by its effectiveness in allowing access to complete phonological representations of speech for deaf people from the age of one month, access to language and eventually performance in reading and writing similar to that of hearing people. Finally with the current high level of development of cochlear implants this method helps facilitate access to the auditory modality.

A large amount of work has been devoted to the effectiveness of Cued Speech but none has investigated the motor organisation of Cued Speech production, i.e. the coarticulation of Cued Speech articulators. Why might the production of an artificial system as long ago as 1967 be of interest? Apart from the clear evidence that such a coding system helps in acquiring another artificial system such as reading, Cued Speech provides a unique opportunity to study lip-hand coordination at syllable level. This contribution presents a study of the temporal organisation of the manual cue in relation to the movement of the lips and the acoustic indices of the corresponding speech sound, in order to characterise the nature of the syllabic structure of Cued Speech with reference to speech coarticulation.

1.2 Overview on Manual Cueing

1.2.1 Cued Speech System

Cued Speech was designed to complement speech-reading . Developed by Cornett (Cornett 1967; Cornett 1982), the system is based on the association of lip shapes to cues formed by the hand. While uttering the speaker uses one hand to point out specific positions around the mouth, palm towards the speaker so that the speech-reader can see the back of the hand simultaneously with the lips. The cues are formed along two parameters: hand placement and hand shape. Positions of the hand code vowels while hand shapes (or configurations) distinguish the consonants. In English, eight hand shapes and four hand placements are used to group phonemes (Figure 1-1). The primary factor in assignment of phonemes to groups associated with a single hand shape or hand position is the visual contrast at the lips (Woodward and Barber 1960). For example, phonemes [p], [b] and [m], with identical visual shapes, are linked to different hand shapes while phonemes easily discriminated from the lips alone are grouped in a single configuration. Each group of consonants is assigned to a hand shape. For the highest frequency group the hand shapes that require less energy to execute are chosen. The frequency

[1]. In memory of Orin Cornett who invented the Cued Speech method at Gallaudet University and to Christian Benoît who initiated Cued Speech synthesis at the ICP laboratory (France).

of appearance of consonant clusters and the difficulties these might present in changing quickly from one hand configuration to another are also taken into account.

Vowel grouping was worked out similarly with high priority being given to the ease of cueing for diphthongs. Vowel positions are indicated with one of the fingers. The middle finger is used for all the consonant cues except those of the [d, p, ʒ], [j, tʃ] and [l, ʃ, w] groups, for which the index finger is used. An exception exists for the [j, tʃ] group: The middle finger is used as the pointer for the mouth position, while the index finger is used for the chin, larynx and side positions.

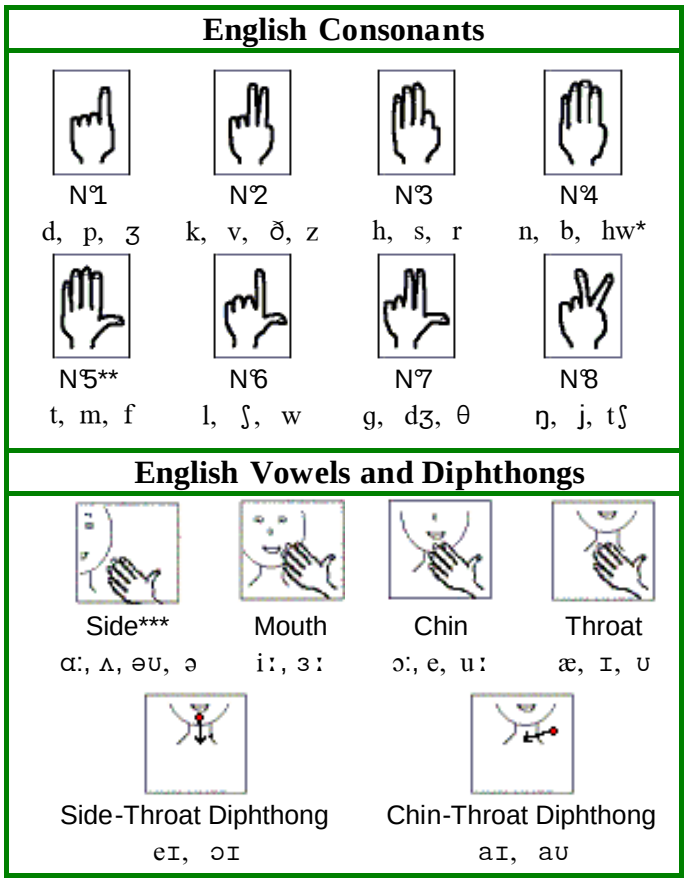

Figure 1-1. Visible cues for English consonants, vowels, and diphthongs (from Cornett 1967). Notes: * Some teachers of Cued Speech may prefer to cue /hw/ as /h/ plus w; ** This hand shape is also used for a vowel without a preceding consonant; *** The side position is used also when a consonant is cued without a following vowel.

The information given by the hand is not sufficient for phoneme identification. The visible information of the lips is still essential. The identification by the lips of a group of look-alike consonants and the simultaneous identification of a group of consonants by the hand shape result in the identification of a single consonant. Thus the combination of hand shape and hand location with the information visible on the lips identifies a single consonant-vowel syllable.

The system was based on the CV syllabification of speech. The syllable strings $C(C_n)V(C_m)$, as complex as they can be, are broken down into CVs each CV being coded both by the shape of the hand for the consonant and by the place of the hand on the face side for the vowel. When a syllable consists only of a vowel, this V syllable is coded using hand shape N°5 (Figure 1-1), with the hand at the appropriate position for the vowel. If a consonant cannot be linked to a vowel, as is the case when two consonants follow each other or when a consonant is followed by a schwa, the hand is placed at the side position with the associated consonant hand shape. Diphthongs are considered to be pairs of

vowels (VV) and are therefore cued with a shift from the position of the first vowel towards the position of the second vowel (cf. Figure 1-1).

Finally in the adaptation of Cued Speech to other languages (more than 50 in Cornett 1988), the criterion of compatibility with the English version was given a higher priority than phoneme frequency of the considered language. An additional position next to the cheekbone is needed for coding all vowels used in French, German, Italian and Spanish. In German some hand shapes code consonant clusters directly (as it is the case for the frequently encountered [ʃt], [ʃp], [tʃ] and [ʃv] clusters) to avoid affecting speech rhythm, a problem that would occur with frequent consecutive hand shape modifications (Pierre Lutz, personal communication).

1.2.2 Perceptual Effectiveness of Manual Cueing

The perceptual effectiveness of Cued Speech has been evaluated in many studies. Nicholls & Ling (1982) presented 18 profoundly hearing-impaired children with CV or VC syllables made up of 28 English consonants together with the vowels [i, a, u] in seven conditions with auditory, lip-reading and manual cue presentations combined. A similar test was conducted with familiar monosyllabic nouns inserted in sentences. Under audition (A) alone subjects correctly identified 2.3% syllables, while scores in lip-reading (L), audition + lip-reading (AL), manual cues alone (C) and audition + manual cues (AC) reached 30 to 39% without significant differences. Higher scores were obtained with lip-reading + manual cues (LC = 83.5%) and audition + lip-reading + manual cues (ALC = 80.4%). This latter result was also recorded for the test sentences where the mean scores for key words reached more than 90% in the LC and ALC conditions.

Uchanski et al. (1994) confirmed the effectiveness of Cued Speech for the identification of various conversational materials (sentences with high or low predictability). The highly trained subjects obtained mean scores varying from 78% to 97% with Cued Speech against 21% to 62% with lip-reading alone.

In French Alégria et al (1992) tested deaf children who had been exposed to Cued speech early (before the age of three) both at home and at school. They compared these early-exposed children with children exposed late from the age of six and only at school. The subjects exposed early and intensively to Cued Speech were better lip-readers and better Cued Speech readers in identifying words and pseudo words. It seems that early exposure to Cued Speech allows children to develop more accurate phonological representations (Leybaert 2000). Thereafter their reading and writing skills progress in a similar way to those of hearing children since Cued Speech early-exposed deaf children can use precise grapheme to phoneme correspondences (Leybaert 1996).

Finally the studies on working memory of Cued Speech deaf children reveal that they use a phonological loop probably based on the visual components of Cued Speech: mouth shapes, hand shapes and hand placements (Leybaert and Lechat 2001).

1.2.3 Phonological Representations in Cued Speech

Fleetwood and Metzger (1998, p. 29) proposed the term *cuem*, which “refers to an articulatory system that employs non-manual signals (NMS) found on the mouth and the hand shapes and hand placements of Cued Speech to produce visibly discrete symbols that represent phonemic (and tonemic) values”. Neither the production nor the reception of acoustic information or of speech is implied in the term “*cuem*”. The authors maintain that Cued Speech can be delivered without production of an acoustic speech signal. This is the usual situation in an interpreting task where the Cued Speech speaker translates silently into cues for deaf people as the hearing speaker is talking. The authors also refer to the studies of Nicholls (1979) and Nicholls and Ling (1982), which claim that the acoustic signal is not necessary in Cued Speech. Nicholls & Ling (1982) found no advantage in audition for

syllable identification. The score obtained in the Cued Speech presentation (manual cues alone; C = 36%) was not significantly different from the Audition + Cued Speech score (AC = 39%). Similarly, there was no difference between the lip-reading + Cued Speech condition (LC = 83.5%) and audition + lip-reading + Cued Speech condition (ALC = 80.4%). The pattern of results was quite different for key words; a better score was recorded for the AC condition (59.2% for low predictability sentences and 68.8% for high predictability) than for the C condition (respectively, 42.9% and 50.0%); in LC and ALC, key word scores were similar, around 96%, revealing a ceiling effect. The advantage of the AC condition for key words in sentences was explained as the use of supra-segmental information. Nicholls & Ling (1982) concluded that speech information in Cued Speech can be perceived through vision alone. Fleetwood and Metzger (1998) proposed that the phonological representations underlying the perception of Cued Speech be defined only by mouth shapes, hand shapes and hand positions (Fleetwood and Metzger 1998).

However we think this position may be too restrictive. In their taxonomy of tactile speech perception methods, Oerlemans and Blamey (1998) proposed distinguishing between speech-based and language-based tactile codes. A code was considered speech-based when the user had direct access to the articulatory gestures, as in the TADOMA method (Reed, Rabinowitz et al. 1985), where the deaf-blind user directly touches the vocal tract of the speaker, placing a hand on the latter's face. In contrast the tactile version of Sign Language was classified as language-based. If the same taxonomy for visual perception is used speech-based and language-based methods can be distinguished. In our view Cued Speech is clearly a speech-based code since the visual lip and mouth information directly results from the articulatory gestures. The fact that the emission of sound is not necessary for the production or reception of Cued Speech does not mean that the code is purely visual. We maintain that Cued Speech is speech-based in the sense that articulatory gestures are recovered from the visual modality. As we will show, these visual lip cues are highly dependent on the speech flow for their temporal time-course.

1.2.4 Face and Hand Coordination for Cued Speech

The fact that manual cues must be associated with lip shapes to be effective for speech perception reveals a real coordination between hand and mouth. As yet no fundamental study has been devoted to the analysis of the skilled production of Cued Speech gestures, i.e. the temporal organisation existing between lip movements and hand gestures in relation to the acoustic realisation [2]. Except for a theoretical aside by Cornett pointing out some consonant clusters where speech should be delayed to leave the hand enough time to reach the correct position (Cornett 1967, p.9), the problems of cue presentation timing are only incidentally touched on in the course of technological investigations[3].

In the Cornett Autocuer system (Cornett 1988) cues are defined from the sound recognition of the pronounced word and are displayed in groups of LEDs on glasses worn by the speech-reader The

[2]. The first studies on Cued Speech production were conducted at the ICP laboratory by Attina and colleagues from 2001 (Attina, Beautemps et al. 2002; Attina, Beautemps et al. 2002; Attina, Beautemps et al. 2002; Attina, Cathiard et al. 2002; Attina, Beautemps et al. 2003; Attina, Beautemps et al. 2003; Cathiard, Attina et al. 2003; Attina, Beautemps et al. 2004; Attina, Beautemps et al. 2004) in the context of a French CNRS “Jeune Equipe” project and a French Research Ministry Cognitive programme. A first prototype of an image-based Cued Speech synthesiser integrating temporal rules has also been realised (Attina, Beautemps et al. 2003; Attina, Beautemps et al. 2004). A 3D model of Cued Speech gestures followed (Gibert, Bailly et al. 2004; Gibert, Bailly et al. 2004; Gibert, Bailly et al. 2004; Gibert, Bailly et al. 2005).

[3]. “When two consonants precede a vowel, as in the word *steep*, the first consonant is cued in the base [side] position and the hand moves quickly to the vowel position while the second consonant cue is formed, in synchronisation with the lip movements. The lips should assume the position for the first consonant as it is cued, but one should not begin making the sound until the hand is approaching the position in which the contiguous consonant and the following vowel are to be cued. This makes it possible to pronounce the syllable naturally." This instruction clearly means that the cuer should wait until the covering [i] vowel gesture has settled before beginning to utter the [s], which is artificially coded with a schwa instead of its natural [i] covering.

whole process involves a delay of 150 to 200 ms for the cue display, compared to the production time of the corresponding sound. This system, designed for isolated words, attained 82% correct identification.

In the system for the automatic generation of Cued Speech developed by Duchnowski et al. (2000) for American English the cues are presented with the help of pre-recorded hands, and rules for temporal coordination with sound are proposed. This system uses a phonetic recogniser of audio speech to obtain a list of phones which are then converted to a time-marked stream of cue codes. The appropriate cues are visually displayed by superimposing hand shapes on the video signal of the speaker's face. The display is presented with a delay of two seconds, a delay that is necessary to correctly identify the cue (since the cue can only be determined at the end of each CV syllable). The superimposed hand shapes are always digitised images of a real hand. Scores of correct word identification reached a mean value of 66% and were higher than the 35% obtained with speech-reading alone but they 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. In this "synchronous" display, the time at which cues were displayed was advanced by 100 ms relative to the start time determined by the recogniser; i.e. for stop consonants, the detected instant of acoustic silence (Duchnowski, personal communication). This advance was fixed empirically by the authors.

In these investigations, the time of cue presentation is related only to the corresponding acoustic events: there is no discussion of the relation between cue presentation and lip motion. However it is well known that lip gesture can anticipate acoustic realisation (Perkell 1990; Abry, Lallouache et al 1996 for French). In the Autocuer system the cue presentation is automatically later than lip motion. The impact of this delay was not evaluated and the identification scores were still high for isolated words. On the other hand, the closer timing of the hand to the acoustic realisation is a key factor in the improvement of the Duchnowski et al (2000) system. It should be stressed that this later system functions with continuous speech and uses hand cues thus it is closer to Manual Cued Speech conditions than the Autocuer .

1.3 First Results on Cued Speech Production

It has been mentioned that the Cued Speech system is based on CV syllabic organisation, the hand giving information on both the consonant and the vowel. The shifting of the hand between two hand positions corresponds to the vocalic transition and the hand shape (or finger configuration) constitutes the consonant information. The main objective of this section is to determine precisely how the hand gesture *co-produces* the consonantal and vocalic information. In short, is the temporal organisation of vocalic and consonant hand gestures similar to the organisation of speech, as revealed by the classical model of coarticulation (Öhman 1967)?

To this end we will examine a comparative study of the temporal organisation of manual cues with lip and acoustic gestures. The temporal organisation of Cued Speech articulators is analysed from a recording of a Cued Speech speaker. The time course of the lip parameter and the hand x y co-ordinates are investigated in relation to acoustic events. The occurrence of hand shape formation is measured in relation to hand position.

1.3.1 The Cued Speech Speaker

The Cued Speech speaker is a 36-year-old French female who has been using Cued Speech at home with her hearing-impaired child for eight years. She qualified in Cued Speech for French in 1996 and regularly translates into Cued Speech code at school.

1.3.2 Audiovisual Data

The different parameters involved in the analysis were derived from the processing of an audio-visual recording of the Cued Speech speaker. The recording was made in a soundproof booth, at 50 frames/second. A first camera in wide focus was used for the hand and the face. A second one in zoom mode dedicated to the lips was synchronised with the first one. The lips were made up in blue. Coloured marks were placed on the hand for tracking hand movement. A second experiment was devoted to the analysis of hand shape formation. In this investigation the Cued Speech speaker was wearing a data glove with two sensors for each of the five fingers covering the first and second articulation with an additional sensor between the fingers. The sensor raw data ear has a linear relationship to the deviation angle between two segments of a finger articulation. The hand position is located with the use of coloured marks placed on the glove. In both experiments, the subject wore opaque goggles to protect her eyes against the halogen spotlight and her head was maintained in a fixed position with a helmet. Blue marks were placed on the speaker's goggles as reference points.

Two Betacam recorders had to be synchronised. At the beginning of the recording session a push button was activated, switching on the set of LEDs (placed in the field of the two cameras) during the first A-frame instant of the video image. This enabled the correspondence between the time codes of the two cameras be calculated. The audio line was digitised in synchrony with the video image. When the data glove was used a system for synchronisation with the audio part was needed. In this system an audio signal was released at the thumb and index finger contact and recorded on the audio line of the video tape. Finger contact resulted in a plateau on the raw data from the glove sensors measuring the movement of the two fingers which allowed synchronisation of the data glove with the audio recording. The delay between the time codes of the two cameras was calculated using the first system.

The image processing-based automatic extraction system developed at ICP (Lallouache 1991) provided a set of lip parameters every 20 ms. We chose to explore the temporal evolution of the between-lip area (S), which is a good parameter for characterising sounds at both the acoustic and articulatory levels. In synchrony with lip area parameter and audio signal, the x and y co-ordinates of the hand mark placed near the wrist were recorded. The beginning and end of hand and lip gesture transitions were manually labelled at the acceleration peaks[4] (Schmidt 1988; Perkell 1990). On the audio signal the beginning and end of the acoustic realisation for consonants and the vowels were also labelled.

These two experiments had complementary objectives. The first explored the movement of the hand from one hand position to another, i.e. the carrier gesture of Cued Speech. Because the hand shape was fixed, interference with hand shape formation was avoided. The second experiment tested the timing of the production of hand shape formation in relation to hand position.

[4]. Velocity and acceleration profiles are derived at each instant from first and second order development of the 4 Hz low pass filtered position, respectively.

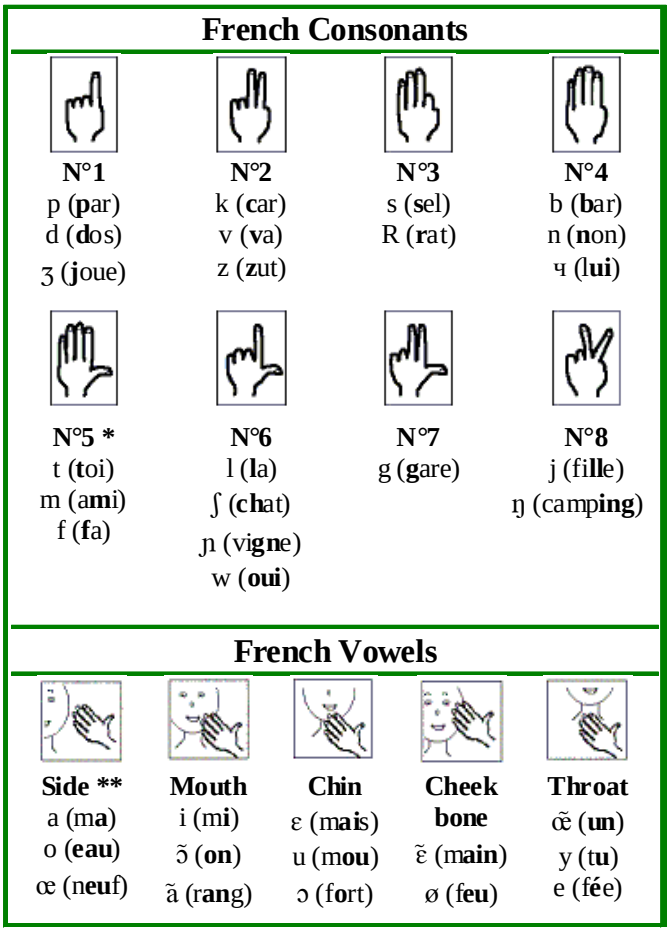

Figure 1-2. Hand positions and hand shapes used in French. Notes: * This hand shape is also used for a vowel not preceded by a consonant. ** This position is also used when a consonant is isolated or followed by a schwa.

1.3.3 Experiment 1: Hand Displacement

1.3.3.1 Corpus

Displacement of the hand was analyzed with [$CaCV_1CV_2CV_1$] sequences made up of [m, p, t] consonants for C combined with the vowels [a, i, u, ø, e] for $V_1$and V_2, i.e., the vowel with the best visibility for each of the five hand positions of the French code (Figure 1-2).

The choice of consonants was fixed according to their labial or acoustic characteristics: [m, p] present a typical bilabial occlusion that appears on the lip video signal as a null lip area, and [p, t] are marked by a clear silent period. The hand shape was fixed during the production of the whole sequence: [m] and [t] are coded with the same hand shape as isolated vowels are (hand shape N°5), while [p] is associated with hand shape N°1. The whole corpus contained twenty sequences, such as [mamamima], for each of the three consonants. A control condition with no consonant for the second (S_2) and third (S_3) syllables was also used, i.e., [$maV_1V_2mV_1$], made up of the vowels [a, i, u, ø, e] for V_1 and V_2 (e.g., [maaima]). We thus obtained twenty additional sequences. For each of the eighty sequences the analysis was carried out on [CV_2] or [V_2] in the absence of a consonant (i.e. on transitions from the S_2 syllable toward S_3 and from S_3 toward S_4), in order to avoid the biases inherent at the beginning of the gesture.

Consider for example the [pupøpu] $S_2S_3S_4$ sequence (from the whole [papupøpu] $S_1S_2S_3S_4$ sequence) in Figure 1-3. The following events were determined for the hand trajectory:

- M1 is the beginning of the hand gesture (determined by acceleration peak) towards the position corresponding to S_3;

- M2 is the hand position target reached (coding S_3), it is determined by peak deceleration and maintained until M3 the instant of peak acceleration and the time at which the hand begins the gesture towards the following position for S_4 coding;
- M4 corresponds to the S_4 hand target reached. In the case of non-concordance of acceleration events on x and y, the first M1 and M3 and the last M2 and M4 points were considered. The hand target is defined as a time when the hand reaches the target both in x and y, i.e. between the end of the transition and the beginning of the transition towards the following target.
- For lip area, L1 marks the beginning of the vowel gesture. This was easily detectable for sequences with [p] and [m] consonants, since L1 was coincident with the end of the lip closure phase. We used the beginning of the acoustical silence to determine L1 in the case of sequences with [t]. L2 is the lip target instant labelled at the end of the lip transition towards the maximal lip-opening target (in the case of absence of a lip vocalic plateau the acceleration peak coincided with the maximal lip value).
- For the corresponding acoustic signal A1 marks the beginning of the consonant of the S_3 syllable.

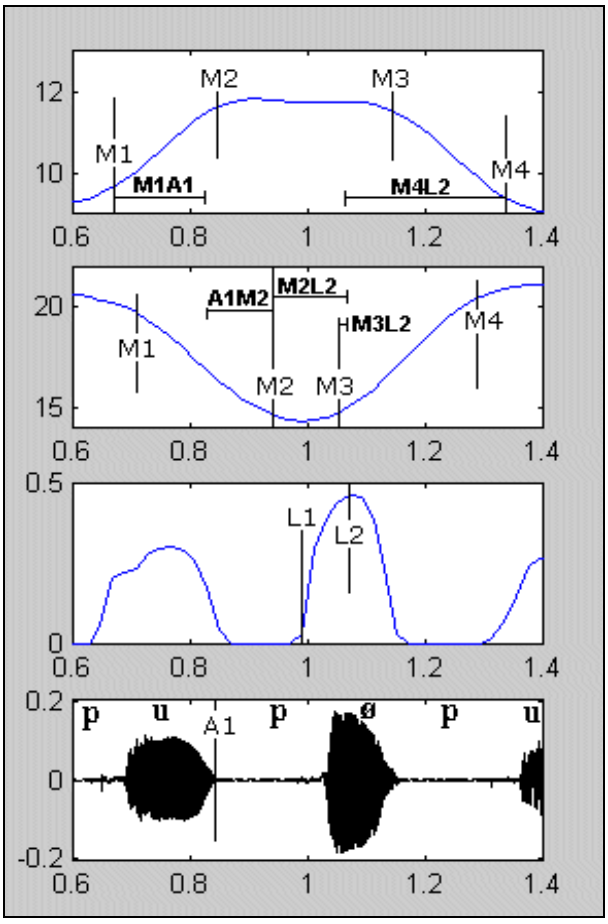

Figure 1-3. From top to bottom: horizontal x (cm) and vertical y (cm) hand motion paths are shown in the top two frames (an increase in x means that the hand moves from the face to the right side, an increase in y means the hand moves towards the bottom of the face). The two bottom frames contain the lip area (cm^2) time course and the corresponding audio signal for the [pupøpu] sequence.

1.3.3.2 Results

For this analysis we took into account only the transitions from the S_2 syllable toward S_3 and from S_3 towards S_4. In order to evaluate the coordination between lip, hand, and sound, we determined

different duration intervals. From the events labelled on each signal, we located the following intervals:

M1A1 corresponds to the interval between the beginning of the manual gesture for S_3 and the acoustic consonant closure.

A1M2 is the interval between the acoustic consonant closure and the onset of the hand target.

M2L2 is the interval between the onset of the hand target and the onset of the lip target of the vowel of S_3.

M3L2 is the interval between the lip target and the beginning of the following hand Cued Speech gesture.

All intervals were computed as arithmetic differences, i.e. the second label minus the first. For example, M1A1 = A1 – M1 (ms). For sequences without a consonant in S_2S_3, such as [maaima], mean values of 183 ms were obtained for the M1A1 interval and 84 ms for the A1M2 interval, the A1 instant corresponding to the onset of the glottal stop that the speaker inserted between the production of the two consecutive vowels. The hand target is clearly in advance of the lip area target (M2L2 = 73 ms). The following hand gesture begins after the lip target (M3L2 = -84 ms).

For sequences with consonants, such as [mamamima], a mean value of 239 ms was obtained for the M1A1 interval. This differed significantly from the consonant acoustical beginning. The A1M2 interval reached a mean value of 37 ms. The hand target was therefore reached during the acoustic realisation of the consonant in a quasi-synchronization with the acoustic closure event. The lip target was usually reached after the corresponding hand target since a mean value of 256 ms for M2L2 was obtained. Finally the hand movement towards the following syllable placement began, on average 51 ms before the peak of the vowel lip target (M3L2 = 51 ms).

In conclusion, the hand gesture begins before the acoustical onset of the CV syllable (183 ms and 239 ms) and usually reaches the hand position well before the lip target, in fact during the consonant.

Figure 1-4. Cues for the [mabuma] sequence.

1.3.4 Experiment 2: Hand Shape Formation

This experiment examined the association between hand shape formation and consonant information. The corpus was selected so as to have only one finger component per consonant hand shape transition in each sequence. For example, the transition from [p] to [k], i.e., from hand shape N°1 to hand shape N°2 (Figure 1-2) is effected by the extension of the middle finger. Thus the modification of the hand shape required only one main sensor of the data glove. This choice was made to simplify data reading.

1.3.4.1 Corpus

Hand shape formation was analysed for two kinds of sequences:

(i) [mVC_1VC_2V] sequences with the same vowel (V = [a] or [ɛ]) were designed to investigate consonant variation. The C_1 and C_2 consonants were [p] and [k], [s] and [b] or [b] and [m]. This choice resulted in hand shape modification at fixed hand placement (for example, the [mapaka] sequence is coded at the side position with the appropriate hand shape modifications). Ten repetitions of each sequence were recorded. The analysis focused on the C_1V syllable, resulting in 60 syllables (10 repetitions x 3 consonant groups x 2 vowels).

(ii) [m$V_1C_1V_2C_2V_1$] sequences varied both vowel and consonant, thus involving both hand shape modification and hand placement transitions. The C_1 and C_2 consonants were [p] and [k], [ʃ] and [g], [s] and [b], or [b] and [m]. The V1 and V2 vowels were [a] and [u], [a] and [e], or [u] and

[e]. Thus, for the [mabuma] sequence (see Figure 1-4) coding implicates a transition of the hand from the side position towards the chin and then back to the side position, while the hand shape changes from the N°5 to N°4 configuration and back to the N°5. The change from 5 to 4 is realised with the thumb facing towards the palm. Five repetitions of each sequence were recorded. The analysis focused on the C_1V_2 syllable, resulting in 60 syllables (5 repetitions x 4 consonant groups x 3 vowel groups). Since an error occurred in the recording for a realisation of a [mubemu] sequence, 59 sequences were considered for this corpus.

In all sequences (with vowel-not-changed and vowel-changed), the beginning of the consonant (A1) is labelled on the acoustic signal. The beginning of the finger gesture is marked at the D1 maximum point of acceleration and the end is marked at the D2 deceleration point of the corresponding raw data trajectory. Similarly for sequences with hand movement from a hand position to another (case of vowel-changed sequences), the hand trajectory was marked by M1 and M2 (Figure 1-5).

1.3.4.2 Results

It should be remembered that the analysis focused only on the second syllable. In order to evaluate the coordination between sound, finger, and hand different duration intervals were derived from the events labelled on each signal. For all the sequences:

D1A1 is the interval between the beginning of the finger gesture and the beginning of the corresponding acoustic consonant;

A1D2 corresponds to the interval between the beginning of the acoustic consonant and the end of the digit movement.

In addition, for vowel-changed sequences:

M1A1 is the interval between the beginning of the hand movement and the beginning of the acoustic consonant;

A1M2 corresponds to the interval between the acoustic consonantal beginning and the end of the hand gesture.

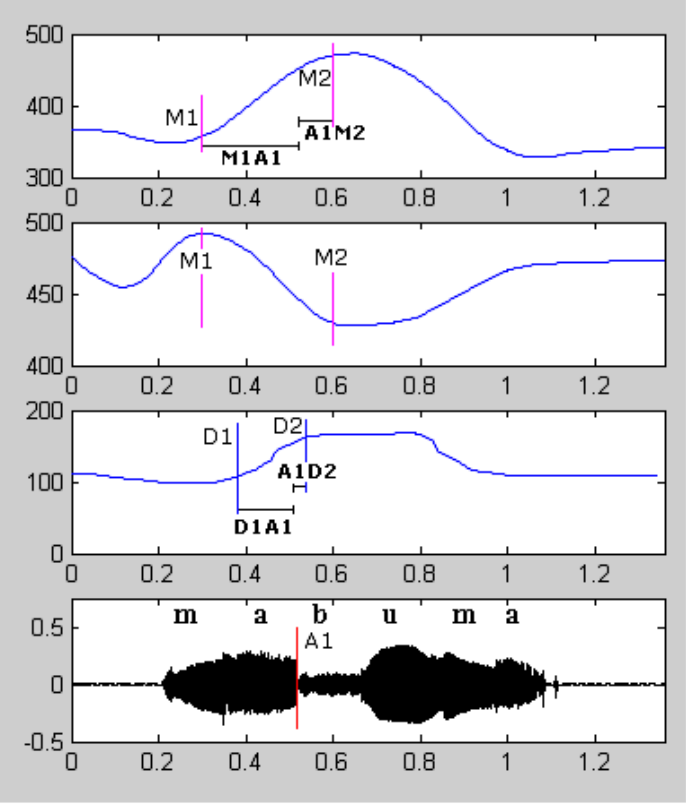

Figure 1-5. From top to bottom: Horizontal x (cm) and vertical y (cm) hand motion paths are shown in the top two frames (An increase in x means moving the hand from the face to the right side, an increase in y means moving the hand towards the bottom of the face); the bottom two frames contain the temporal deviation of the raw data of the thumb first articulation glove sensor and the corresponding acoustic signal for the [mabuma] sequence.

As in the first experiment, all intervals were computed as arithmetic differences, i.e. the second label minus the first label, for example, D1A1 = A1 – D1 (ms).

For the vowel-not-changed sequences (sequences with only hand shape change, the hand placement being maintained), we obtained mean values of 124 ms for the D1A1 interval and 46.5 ms for the A1D2 interval. Thus the beginning of the finger gesture precedes the acoustic beginning of the consonant. The finger finishes its movement just after the beginning of the acoustic realisation of the consonant.

For the vowel-changed sequences (both hand shape and hand placement change), mean values of 171 ms for the D1A1 interval and -3 ms for the A1D2 interval were obtained. Thus, for the finger gesture relative to the sound, we observed the same pattern as in the previous result. As regards the hand gesture, mean values of 205 ms for the M1A1 interval and 33 ms for the A1M2 interval were obtained. The hand gesture begins before the finger gesture and consequently well before the onset of the acoustic consonant. The hand target is reached at the beginning of the acoustic realisation of the consonant. Finally if we compare duration for hand shape formation in reference to hand transition between two hand placements, we note that the consonant finger gesture is encapsulated in the hand transition.

1.3.5 Summary of the Two Experiments

There is a noticeable convergence in the results of the two experiments. To summarise, for hand position it was observed that

- the movement of the hand towards its position begins about 200 ms before the acoustic beginning of the CV syllable. This implies that the gesture begins during the preceding syllable, i.e. during the preceding vowel;
- the hand target is attained at the beginning of the acoustic consonant beginning;
- this hand target is therefore reached on average 250 ms before the vowel lip target.

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

Finally, it was observed from the data glove that the hand shape is completely formed at the instant when the hand target position is reached. In addition it was noticed that the hand shape formation gesture uses a large part of the hand transition duration.

1.4 General Discussion

1.4.1 Cued Speech Co-production

The consideration of the two Cued Speech components within the framework of speech control has a bearing on the future elaboration of a quantitative control model for Cued Speech production. For transmitting consonant information, the control type is figural, i.e., a postural control of the hand configuration (finger configuration). The type of control for transmitting the vowel information is a goal-directed movement performed by the wrist and 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:

(i) The mandibular open-close oscillation is the control of a cycle, self-initiated and self-paced (Mac Neilage, 1998; Abry, Stefanuto, Vilain & Laboissière, 2002). This is the control of the carrier of speech, the *proximal* control that produces the syllabic rhythm.

(ii) Following Öhman (1967; see also Vilain, Abry & Badin, 2000), the vowel gesture is produced by *global* control of the whole vocal tract – from the glottis to the lips --, i.e. a figural or postural motor control type.

(iii) The consonant gesture is produced by the control of contact and pressure performed *locally* along the vocal tract.

The carried articulators (tongue and lower lip) together with their coordinated partners (upper lip, velum and larynx) are involved in these two distal (global and local) controls.

The mandibular and vowel controls are coupled by in-phase locking. Consonantal control is typically in phase with the vowel for the initial consonant of a CV syllable but it can be out-of-phase for the coda consonant in a CVC syllable. Finally consonant 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 regards speech, Cued Speech vowels and consonants 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 local placement of the hand around the face. On the other hand, the consonant *carried* gesture is a postural (figural) one. Thus the two types of control in Cued Speech are inversely distributed in comparison to speech: the configuration of global control of the speech vowel corresponds to a local control in Cued Speech whereas the local control for the speech consonant corresponds to a global control in Cued Speech.

Once speech rhythm has been converted into Cued Speech rhythm (that is a general CV syllabification with some cluster specificities as in German), the two carriers (mandible and wrist) can be examined with respect to their temporal coordination, i.e. phasing. This CV re-syllabification means that every consonantal Cued Speech gesture will be in phase with its vocalic one, which is not always the case in speech for languages that have more than just CVs. Unlike speech the Cued Speech consonant gesture never hides the beginning of the in-phase vocalic gesture (Öhman's model). As for the phasing of the two carried vowel gestures, our experiments made clear that the Cued Speech vowel gesture did anticipate the speech vowel gesture.

1.4.2 Towards an Upside Down Vision of Cued Speech

The coordination obtained between hand, lips, and sound confirms, in our opinion, the in-principle validity of the advance (lead) of the hand on the sound, programmed as an empirical rule by Duchnowski et al. (2000) for their automatic Cued Speech display. Obviously the range of this anticipatory behavior will vary with different speakers, rates etc and should be examined by subsequent articulatory studies.

These considerations result in quite a rather upside down vision of the Cued Speech landscape. The *in-principle* advance of the hand over the lips (and on sound) is crucial for the question of the integration of manual and lip information. Currently Cued Speech has been designed as an augmentation for lip disambiguation. A general pattern seems to appear from our data on the temporal organisation of hand and lip gestures in the production of successive CV sequences. The hand attains the vowel placement at the beginning of the CV syllable and moves from that position towards a new one even before the peak acoustic realisation of the vowel and before the corresponding vocalic lip target is reached. It seems therefore that production control imposes its temporal organisation on the perceptual processing of Cued Speech. This organisation leads us to think that the hand placement first gives a set of possibilities for the vowel then the lips then determine a unique solution. This hypothesis has been successfully tested within the framework of gating experiments for phoneme identification where recognition of CV syllables has been evaluated across the time course of available online

information resulting from the coordination of hand and lip motion (see Cathiard, Bouaouni et al. 2004; Attina, 2005; Troille, Cathiard et al. 2007 and Troille, 2009). These studies demonstrated the ability of deaf subjects Cued Speech users to recover the anticipatory behaviour of the hand in their Cued Speech perception.

1.5 Acknowledgments

Many thanks to Martine Marthouret speech therapist at Grenoble Hospital, for helpful discussions; to Mrs. G. Brunnel, the Cued Speech speaker, for enduring the recording conditions, and C. Abry and J. L. Schwartz for their stimulating suggestions. This work has been 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 the CNRS.

1.6 References

Abry, C., M.-T. Lallouache and M.-A. Cathiard (1996). How can coarticulation models account for speech sensitivity to audio-visual desynchronization? Speechreading by Humans and Machines. D. Stork and M. Hennecke. Berlin, Springer-Verlag**:** 247-255.

Alégria, J., J. Leybaert, B. Charlier and C. Hage (1992). On the origin of phonological representations in the deaf : Hearing lips and hands. Analytic Approaches to Human Cognition. J. Alégria, D. Holender, J. Junça de Morais and M. Radeu. Amsterdam, Elsevier Science Publishers**:** 107-132.

Attina, V., D. Beautemps and M.-A. Cathiard (2002). Contrôle de l'anticipation vocalique d'arrondissement en Langage Parlé Complété. Journées d'Etudes sur la Parole, Nancy - France

Attina, V., D. Beautemps and M.-A. Cathiard (2002). Coordination of hand and orofacial movements for CV sequences in French cued speech. International Conference on Speech and Language Processing, Boulder - USA, 1945-1948.

Attina, V., D. Beautemps and M.-A. Cathiard (2002). Organisation spatio-temporelle main - lèvres - son de séquences CV en Langage Parlé Complété. Journées d'Etudes sur la Parole, Nancy - France

Attina, V., D. Beautemps and M.-A. Cathiard (2003). Temporal organization of French cued speech production. International Conference of Phonetic Sciences, Barcelona, Spain, 1935-1938.

Attina, V., D. Beautemps and M.-A. Cathiard (2004). Cued Speech production: giving a hand to speech acoustics. CFA/DAGA'04, Strasbourg, France, 1143-1144.

Attina, V., D. Beautemps, M.-A. Cathiard and M. Odisio (2003). Towards an audiovisual synthesizer for Cued Speech: rules for CV French syllables. Auditory-Visual Speech Processing, St Jorioz - France, 227-232.

Attina, V., D. Beautemps, M.-A. Cathiard and M. Odisio (2004). "A pilot study of temporal organization in cued speech production of French syllables: rules for a cued speech synthesizer." Speech Communication **44**: 197-214.

Attina, V., M.-A. Cathiard and D. Beautemps (2002). Controling anticipatory behavior for rounding in French cued speech. Proceedings of ICSLP, Denver, Colorado, 1949-1952.

Attina, V. (2005). La Langue Française Parlée Complétée (LPC): Production et Perception. Unpublished PhD manuscript, Institut National Polytechnique, Grenoble, France.

Bernstein, L. E., M. E. Demorest and P. E. Tucker (2000). "Speech perception without hearing." Perception & Psychophysics **62**: 233-252.

Cathiard, M.-A., V. Attina and D. Alloatti (2003). Labial anticipation behavior during speech with and without Cued Speech. International Conference of Phonetic Sciences, Barcelona, Spain, 1939-1942.

Cathiard, M.-A., F. Bouaouni, V. Attina and D. Beautemps (2004). Etude perceptive du décours de l’information manuo-faciale en Langue Française Parlée Complétée. Journées d'Etudes de la Parole, Fès, Maroc, 113-116.

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

Cornett, R. O. (1982). Le Cued Speech. Aides manuelles à la lecture labiale et perspectives d'aides automatiques. F. Destombes. Paris, Centre scientifique IBM-France.

Cornett, R. O. (1988). "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.

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

Fleetwood, E. and M. Metzger (1998). Cued Language structure. An analysis of Cued american English based on linguistic principles. Silver Spring, Maryland, Calliope Press.

Gibert, G., G. Bailly, D. Beautemps, F. Elisei and R. Brun (2005). "Analysis and synthesis of the 3D movements of the head, face and hand of a speaker using cued speech." Journal of Acoustical Society of America **118**(2): 1144-1153.

Gibert, G., G. Bailly and F. Elisei (2004). Audiovisual text-to-cued speech synthesis. 5th Speech Synthesis Workshop, Pittsburgh, PA, 85-90.

Gibert, G., G. Bailly, F. Elisei, D. Beautemps and R. Brun (2004). Audiovisual text-to-cued speech synthesis. Eusipco, Vienna - Austria, 1007-1010.

Gibert, G., G. Bailly, F. Elisei, D. Beautemps and R. Brun (2004). Evaluation of a speech cuer: from motion capture to a concatenative text-to-cued speech system. Language Resources and Evaluation Conference, Lisbon, Portugal, 2123-2126.

Lallouache, M.-T. (1991). Un poste Visage-Parole couleur. Acquisition et traitement automatique des contours des lèvres. PhD Thesis. Institut National Polytechnique Grenoble.

Leybaert, J. (1996). "La lecture chez l'enfant sourd : l'apport du Langage Parlé Complété." Revue Française de Linguistique Appliquée, Paris, AFLA **1**: 81-94.

Leybaert, J. (2000). "Phonology acquired through the eyes and spelling in deaf children." Journal of Experimental Child Psychology **75**: 291-318.

Leybaert, J. and J. Lechat (2001). "Phonological similarity effects in memory for serial order of cued speech." Journal of speech, language and hearing research **44**: 949-963.

Nicholls, G. (1979). Cued Speech and the reception of spoken language. McGill University Montréal: 163 pages.

Nicholls, G. and D. Ling (1982). "Cued Speech and the reception of spoken language." Journal of Speech and Hearing Research **25**: 262-269.

Oerlemans, M. and P. Blamey (1998). Touch and auditory-visual speech perception. Hearing by Eye: Part 2, The Psychology of Speechreading and Auditory-visual Speech. R. Campbell, B. Dodd and D. Burnham. Hillsdale, NJ, Earlbaum**:** 245-281.

Öhman, S. E. G. (1967). Word and sentence intonation: a quantitative model. Stockholm - Sweden, Speech Transmission Laboratory - Department of Speech Communication and Music Acoustics - KTH**:** 20-54.

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.

Perkell, J. S. (1990). Testing theoris of speech production : Implications of some detailed analyses of variable articulatory data. Speech Production and Speech Modelling. W. J. Hardcastle and A. Marchal. London, Kluwer Academic Publishers**:** 263-288.

Reed, C. M., W. M. Rabinowitz, N. I. Durlach and L. D. Braida (1985). "Research on the Tadoma method of speech communication." Journal of the Acoustical Society of America **77**(1): 247-256.

Reisberg, D., J. McLean and A. Goldfield (1987). 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.

Schmidt, R. A. (1988). Motor Control and Learning: A Behavioral Emphasis. Champaign, IL, Human Kinetics Publishers.

Troille, E., M.-A. Cathiard and C. Abry (2007). A perceptual desynchronization study of manual and facial information in French Cued Speech. ICPhS, Saarbrücken, Germany, 291-296.

Troille, E. (2009). De la perception audiovisuelle des flux oro-faciaux en parole à la perception des flux manuo-faciaux en Langue française Parlée Complétée. Adultes et Enfants: Entendants, Aveugles et Sourds. Unpublished PhD manuscript, Stendhal University, Grenoble, France.

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.

Woodward, M. F. and C. G. Barber (1960). "Phoneme perception in lipreading." Journal of Speech and Hearing Research **3**(3): 212-222.