Cued Speech production: giving a hand to speech acoustics

Virginie Attina, Denis Beautemps, Marie-Agnès Cathiard
Institut de la Communication Parlée, UMR CNRS 5009, Grenoble, France
Email: {attina;beautemps;cathiard}@icp.inpg.fr

Introduction

Manual Cued Speech is an effective method used to enhance speech perception for orally educated deaf people [1]. Thanks to this system, a speaker can clarify what he says with the help of hand cues near the face; lip shapes are thus disambiguated by the addition of a manual. A cue is made up of two components: the shape of the hand (finger configuration) and its position around the face. Handshapes are designed to distinguish among consonants and hand positions are devoted to vowel disambiguation (Figure 1). Seeing manual cues associated to lip shapes allows the cue receiver to identify unique speech elements.

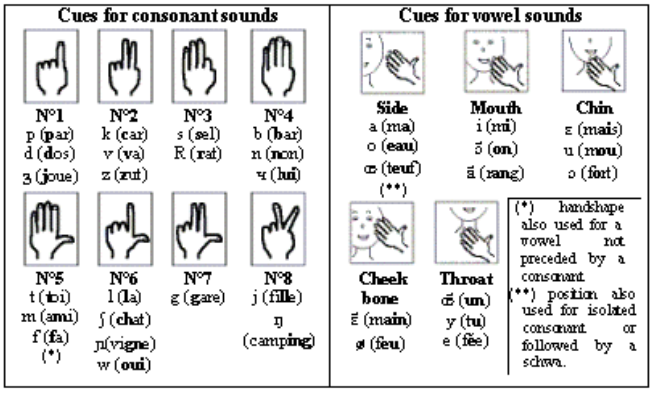

Figure 1: Cues for French vowels and consonants.

A large amount of work has shown the effectiveness of Cued Speech for visual phoneme perception, for complete phonological representations acquisition and then for language acquisition or reading and writing learning (for a review see [2]). 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 contribution outlines 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 French Cued Speech (FCS). A first study conducted on a single FCS transliterator at different periods is presented. Results are replicated for 3 other subjects.

A pilot study on a FCS transliterator

In this section, we present principal results of the temporal coordination of French Cued Speech investigated on a proficient transliterator (GB). For more details on this work, see [3].

Method

Two experiments were conducted. Experiment 1 focused on hand movements with fixed hand shapes in relation with lip gestures and acoustic signal while Experiment 2 investigated how hand shapes are formed relatively to hand transitions, lip gestures and acoustic signal. Sequences made of series of CV syllables were recorded as for example [mamumimu] for Experiment 1 and [mabuma] for Experiment 2.

The speaker GB was audio-visually recorded in a sound-proof booth, at 50 frames/second. Two synchronous cameras were used respectively for the hand in large focus and for the lips in zoom mod. The acoustic signal was digitalized in synchrony with the video and sampled at 22050 Hz. Processing of the video by tracking of colored marks placed on the back of the hand gives the x and y coordinates of the hand marks at a frequency rate of 50 Hz. An automatic extraction of lip contours was processed giving the temporal evolution of lip area at 50 Hz. For the analysis of handshape formation (experiment 2), a data collector glove was added giving raw data of the sensor studied at 64 Hz. Finally the data processing provided synchronous signals. Each signal was labeled with the help of acceleration in order to compare different temporal events between the signals: the beginning and the end of hand transitions (M1 and M2, M3 and M4), the vocalic lip target (L2) and the onset (A1) of consonant acoustic realization for sound.

Results

For experiment 1, a mean duration of 399.5 ms (σ=95.6) for the acoustic realization of CV syllables was found corresponding to a slow mean speech rate of 2.5Hz. The following duration intervals were derived from the labels: (i) M1A1 is the interval between the beginning of the manual gesture and the acoustic consonant onset, (ii) A1M2 is the interval between the acoustic consonant onset and the reach of hand target, (iii) M2L2 is the interval between the reach of the hand target and the vocalic lip target and (iv) M3L2 is the interval between lip target of and the beginning of the following hand FCS gesture coding the following syllable. Each interval's duration was computed as the arithmetic difference (for e.g. M1A1=A1-M1 (ms)). Table 1 shows the different values.

	M1A1	A1M2	M2L2	M3L2
Duration (ms)	239	37	256	51
(std)	(87)	(76)	(101)	(60)
%/CV	60	9	64	13

Table 1: Mean durations and standard deviation for temporal intervals and their corresponding percentage relatively to the CV syllable mean duration

For Experiment II, a mean duration of 316.3 ms (σ=44.6) for the acoustic duration of the syllable was observed corresponding to a mean speech rate of 3.2 Hz. The same intervals as for Experiment 1 were considered: M1A1, A1M2, M2L2 and M3L2. In addition for the handshape: (i) D1A1 is the interval between the beginning of the finger

gesture and the onset of the corresponding acoustic consonant, (ii) A1D2 corresponds to the interval between the onset of the acoustic consonant and the end of the finger movement, (iii) D2L2 is the interval between the finger handshape target and the vocalic lip target and (iv) L2D3 is the interval between the vocalic lip target and the onset of finger movement for the following syllable. Table 2 shows the mean durations calculated for each interval.

	D1A1	A1D2	D2L2	L2D3
Duration (ms) (std)	171 (48)	-3 (45)	208 (91)	53 (72)
%/CV	54	-1	66	17
	M1A1	A1M2	M2L2	M3L2
Duration (ms) (std)	205 (55)	33 (50)	172 (67)	43 (76)
%/CV	65	10	54	14

Table 2: Mean durations for intervals and their corresponding percentages of the CV mean duration

We obtained a noticeable coherence with results of the 2 experiments (Figure 2) involving a same FCS speaker recorded at two different periods after a space of one year. To sum up, concerning hand position, it was observed for a CV syllable: (i) the displacement of the hand toward its position began more than 200 ms before the consonant acoustic onset of the CV syllable, namely that represents 60-65% of the CV duration. This implied that the gesture began in fact during the preceding syllable; (ii) the hand target was attained at the acoustic onset of the consonant that is 9-10% of the whole CV duration; (iii) this hand target was reached largely before the vowel lip target, in average 172 to 256 ms (54-64%) before.

These three results revealed the *anticipatory* gesture of the hand motion over the lips. Finally, it was observed from the data glove that the handshape was completely preshaped at the instant where the hand target position was reached. Moreover we noticed that the handshape formation gesture D1D2 (168 ms) used a large part (71%) of the hand transition duration M1M2 (238 ms).

FCS production by 3 transliterators

Method

Sequences made of CV syllables like [mamabima] were recorded 3 times with different speakers using the same experimental setup as for the previous study. Focus was on hand transitions. The same temporal intervals were calculated.

Results

Results as percentages of the mean CV syllable duration for each transliterator (AM, SC and RV) are presented in Table 3. A closeness of results between the 3 subjects is clearly revealed.

	AM	SC	RV
CV (ms) (std)	252 (41)	253 (45)	256 (57)
Speech rate (Hz)	4	4	3.9
M1A1	61	57	57
A1M2	7	11	18
M2L2	61	56	48

Table 3: For each of the 3 subjects, mean CV durations, corresponding speech rates and temporal intervals calculated as percentages of the mean CV duration

Conclusion

A great coherence of results was found on all the 4 subjects showing up a stable temporal pattern for FCS production. The syllabic speech rates obtained (2.5 to 4 Hz) correspond to the slowing down of the speech usually observed in Cued Speech. Indeed [4] indicates a value of 100 wpm i.e. a range between 3 to 5 Hz for the syllabic rhythm. The anticipatory gesture of hand motion over the lips, well confirmed by the results in 4 proficient transliterators, is revealed as an invariant feature across the subjects. This advance was suggested heuristically by [4]. Considering a CV syllable, the hand begins its movement well before the acoustic onset of the consonant (57 to 65% of the CV duration) and attains its position in the first part of the syllable (7 to 18%), so largely before the vocalic lip target (48 to 64%).

References

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

[2] J. Leybaert and J. Alegria. The Role of Cued Speech in Language Development of Deaf Children, in M. Marschark and P. E. Spencer (Eds), Oxford Handbook of Deaf Studies, Language, and Education (2003), 261-274.

[3] V. Attina, D. Beautemps, M.-A. Cathiard and M. Odisio. Toward an audiovisual synthesizer for Cued Speech: Rules for CV French syllables. Proc. of AVSP, St Jorioz, (2003), 227-232.

[4] P. Duchnowski, L.D. Braida, M. S. Bratakos, D. S. Lum, M. G. Sexton and J. C. Krause. A speechreading aid based on phonetic ASR. Proc. of the 5th ICSLP, Sydney, **7**, (1998), 3289-3292.

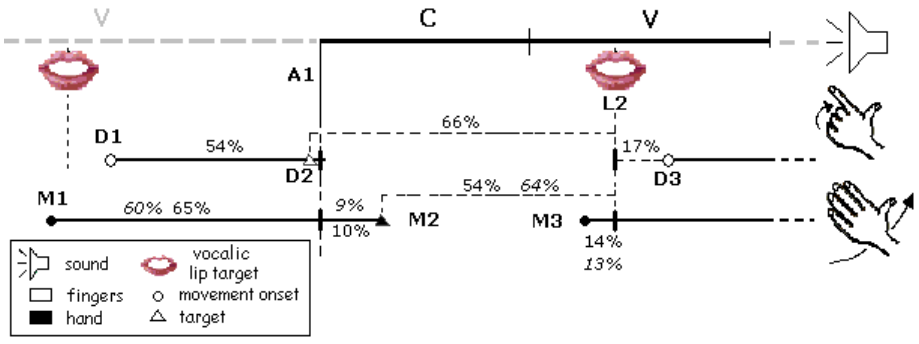

Figure 2: Temporal pattern for coordination between sound, lips, handshape formation and hand placement for FCS production from results of GB. Represented percentages are from CV mean duration (values in italic for Experiment 1).