Two CV syllables for one pointing gesture as an optimal ratio for jaw-arm coordination in a deictic task: A preliminary study

Amélie Rochet-Capellan, Jean-Luc Schwartz, Rafael Laboissière, Arturo Galvan

Two CV syllables for one pointing gesture as an optimal ratio for jaw-arm coordination in a deictic task: a preliminary study

Amélie Rochet-Capellan (amelie.rochet-capellan@icp.inpg.fr)
Jean-Luc Schwartz (jean-luc.schwartz@ic.inpg.fr)
Institut de la Communication Parlée
INPG / Université Stendhal/ CNRS UMR 5009 - INP Grenoble 46, avenue Félix Viallet 38031 GRENOBLE CEDEX 1, France

Rafael Laboissiere (laboissiere@cbs.mpg.de)
Arturo Galvan (galvan@cbs.mpg.de)
Max Planck Institute for Psychological Research Amalienstrasse 33 80799 Munich

Abstract

According to the "Vocalize-to-Localize" framework (Abry et al., 2004), the association of oral and brachiomanual gestures in the deictic function could be the root of the first words emergence in both ontogeny and phylogeny. This association may require a close coordination between the two gestural systems, possibly anchored in a 2:1 harmonic ratio between the natural oscillatory frequencies of the *Speech frame* (open-close jaw cycle) and the Sign frame (arm-hand-finger pointing cycle). This could explain why both languages and first words favor bi-syllabic forms (Ducey-Kaufman et al., 2005). The present study used a new paradigm to test this 2:1 ratio in adult on-line productions. The results provide first evidence that supports the 2:1 ratio and suggest new arguments for a substance-based approach of language evolution.

Introduction

Language and Action

In the last 25 years, language and action have been more and more considered as connected systems. A major evidence for this assumption is the coordination of speech and gestures in on-line face-to-face interactions (McNeill, 1981). Moreover, this coordination appears to be motor rather than purely perceptual. For example, gestures are as well involved in linguistic communication between blind people (Iverson and Goldin-Meadow, 1998). In addition, the link begins early in development, inside the corporal "babbling" of speech and hands (Iverson and Thelen, 1999) and then with the association of words and pointing gestures in the passage from one- to two-words production (Pizzuto et al., 2005; Volterra et al. 2005). The neuroimagery data also lead to a cortical neuroanatomy in which language, perception and action share a same temporo-parieto-frontal circuit (Pulvermüller, 2005, for a review). The Broca's area itself, considered as a "pure language area" for a long time, has been shown to be involved in perception, action understanding and imitation (Nishitani et al. 2005, for a review). Moreover, overlapping brain areas subtend oral and sign languages (Emmorey et al, 2002, MacSweeney et al., 2002) leading to suspect that some characteristics of language are modality-free (San José-Robertson et al., 2004). All this background favors the assumption of a close link between orofacial and brachiomanual actions in the emergence of language.

Deriving language from gestures

Corballis (2003) proposed that language could have first emerged as a manual communication system and then, progressively evolved towards mouth and speech in the course of phylogeny. On the contrary, the "Frame then Content" (FC) theory (MacNeilage, 1998) links the emergence of speech and language to the orofacial motor control system, considering ingestive mechanisms as a primary step towards oral communication. This theory further relates universals in adult languages and infant babbling to an evolutionary-developmental scenario (MacNeilage and Davis, 2000) in which jaw motor cyclicities (*the frame*) would have been primary. Then the independent and coordinated control of the tongue and the lips would have been mastered (*the content*). In this "frame-then-content" sequence experimentally displayed in the course of ontogeny (Munhall and Jones, 1998; Green et al., 2002), the jaw is considered as the carrier of speech gestures. This "frame dominance" would explain the preference for Consonant-Vowel syllable forms in both human languages and infant babbling. However, these "mono-modal" scenarios about language origins gave rise to criticisms (Abry et al., 2004; Arbib, 2005) that let appear a consensus: the two motor systems might have in fact evolved conjointly towards an elaborated communication system involving both brachiomanual and orofacial gestures. Coming back to McNeill, the evolutionary process would have selected the capacity to associate speech and gestures in a coherent communicative process.

Deriving language from deixis

In this evolutionary process, the deictic function might have played a pivotal role. Indeed, the deictic gesture has been considered as the primary indexical sign in both phylogeny and ontogeny (Haviland, 2000). Furthermore, according to Abry et al. (2004) the association of voice and hand in a deictic function is a key step of both the phylogenetic and ontogenetic development of language. They assumed that language would provide a new deictic tool enabling humans to "vocalize to localize". In this "Vocalize to Localize" framework, the baby's – and by extend, the humanity's - first words would be the

product of a developmental "rendez-vous" between the motor control mastery of arm-hand-finger pointing (the *Sign Frame*) and the jaw opening-closing gestures that subtend babbling (the *Speech Frame*, referring to the FC theory). This "rendez-vous" requires the coordination of the two systems, constrained and shaped by the physical and motor properties of the speech and the arm-hand systems. Indeed, Ducey-Kaufmann et al. (2005) displayed a 2:1 harmonic ratio between the *Speech Frame* and the *Sign Frame* for 6 French children between 6- and 18-months old. In other words, the babies tend to utter two CV syllables (associated to two jaw cycles) inside one pointing gesture. According to Ducey-Kaufmann et al., since the first words would emerge from the rendez-vous between the *Speech Frame* and the *Sign Frame*, the 2:1 ratio could explain why both infants' first words and human languages favor two-syllables words (Rousset, 2004). In this "Vocalize to Localize" framework, the goal of this paper is to provide a new experimental evidence for the 2:1 ratio in adult deictic tasks.

Evaluating the 2:1 hypothesis in adults

Previous studies showed a coordination between speech and pointing gestures, mainly resulting from a speech adaptation (Levelt et al., 1985 Feyreisen, 1997). Other studies displayed preferential synergies between speech and finger taping motion (Kelso et al. 1983, Treffner et Peter, 2002). Jaw preferential oscillatory frequency has also been estimated through dynamics studies (Nelson et al., 1984). Yet, no study has tried to establish the favored ratio between the *Speech Frame* and the *Sign Frame* in adult. Here, we propose an original experimental paradigm testing the 2:1 hypothesis for arm-jaw coupling in a deictic task involving utterances with 1, 2, 3 or 4 CV syllables. The basic assumption is that at most two jaw cycles can be contained inside one pointing cycle. Considering that one CV syllable requires one jaw cycle, the pointing cycle should stay constant for 1- and 2- CV syllables utterances. It should then increase for 3-syllables and remain the same for 3- and 4-syllables utterances. This portrait (1=2<3=4) is the focus of the present experiment. In order to measure the period of the arm-finger pointing cycle while avoiding rhythmic tasks and keeping a relatively natural deictic gesture, the task was to show a target while naming it (with 1-, 2-, 3- or 4- CV syllable(s) logatoms) twice in rapid succession.

Method

Procedure

The subjects were 9 native Brazilian female speakers, all right-handed and without any speech or hearing problems. As in the princeps study by Levelt et al. (1985), the experiment involved a gesture + speech pointing task. The subject was seated at a table. A target (red smiley ☺ icon) together with the logatom to pronounce were projected in front of her on a board during 2.5 sec +/- a random delay (with a 1-sec mean and a 0.15-sec standard deviation). The target appeared in the right visual field, either at a near or a far position. The logatom was projected at the same time in the middle of the visual field, as displayed on Figure 1. The logatom could be either /pa/, /papa/, /papapa/ or /papapapa/. It was introduced as a person's name and the target-smiley as a symbolic representation of that person. The instruction was to name and show "the person" two successive times as soon as the icon color changed (go-signal). The subject was invited to put her finger on a black square mark on the table before and after each trial. For, example, for a /pa/ item, the speaker showed the target a first time saying /pa/, put her finger back on the black square and immediately showed the target again saying /pa/. As some subjects tended to confuse /papapa/ with /papapapa/, the experimenter read the logatom aloud in order to remove the ambiguity. This methodology problem will be discussed here after. Before starting, the subject was asked to show objects in the room while naming them in order to link the task with real-world pointing situations. The experimental phase was divided into four blocks separated by a 30-sec pause. Each block started with 4 practice trials followed by 40 experimental ones, five for each of the eight experimental conditions (number of syllables) * (target position). The order of the trials was differently randomized for each block and each speaker.

Data recording

Jaw and hand motions were recorded using an optotrack system at a 100 Hz sample frequency. Two sensors were pasted on the subject's right forefinger, one on the middle of the nail and the other on the left part next to the nail. This allowed keeping the finger in the visible optotrack field when the finger turned towards the right during the pointing gesture. A third sensor was pasted between the top of the chin

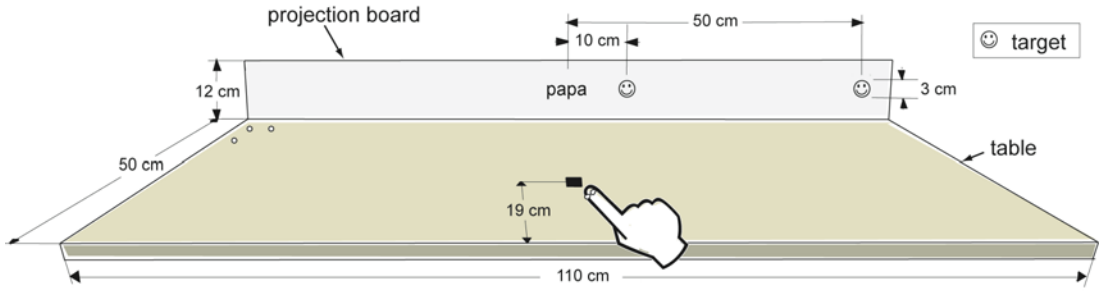

Figure 1: Experimental apparatus (inspired from Levelt et al. (1985))

and the lower lip on the speaker's face. It tracked a flesh point rather than the jaw itself, but in light of the phonetic material used here, with only opening-closing movements between a stop labial consonant and an open /a/, this sensor was considered as a good indicator of jaw motion. For simplification, it is now referred as the jaw sensor. Three fixed sensors on the table and three others maintained on the subject's head provided two referentials, respectively for the finger and jaw moving sensors. The sound was recorded at a 16 kHz sampling frequency with a computer connected to a microphone. Synchronization between acoustic and optotrack signals was achieved by a beep generated on the second sound channel when the optotrack record started.

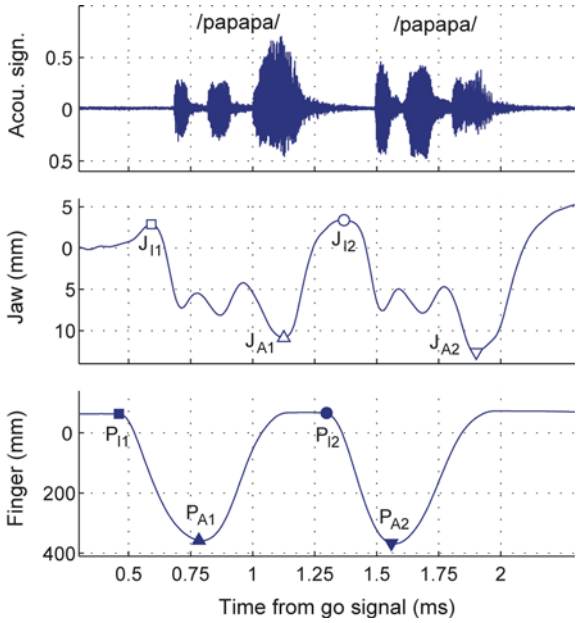

Figure 2: Acoustic signal (top) and trajectories of jaw (middle) and finger (bottom) for a /papapa/ trial. The points marked on the trajectories curves are the initiations and the arrivals of motions labeled for the jaw (J_{I1}, J_{A1}, J_{I2} and J_{A2}) and for the finger (P_{I1}, P_{A1}, P_{I2} and P_{A2}), see text for detail.

Data processing

After the experiment, each trial was checked in order to detect speech errors (e.g. non respect of the required number of syllables) and gesture errors (e.g. departure before the go signal). Then, finger, head and jaw sensors coordinates were projected into the table referential and then, jaw coordinates into the head referential. The trajectory of the three sensors against time was estimated running a Principal Component Analysis on x-y-z sensors coordinates. The first component explained most of the variance for all subjects (from 92% for the jaw to 98% for both finger sensors) and so, was used as the estimator of sensors trajectories against time. These signals were lowpass filtered by a Butterworth filter with a cutting frequency at 15 Hz. On these valid trials signals, onset and offset events were positioned on finger and jaw trajectories (respectively for speed increasing above or decreasing under a threshold, set at 10% of the maximum speed on the corresponding stroke). Thus, P_{I1}, P_{I2} and P_{A1}, P_{A2} are, respectively, the initiation (onset) and the apex (offset) times from the go signal for the first and the second finger pointing gesture (Figure 2, bottom). Similarly, J_{I1} and J_{I2} are the initiation (onset) events of the jaw opening motion for the first syllable of the two utterances. Finally, J_{A1} and J_{A2} are the apex (offset) events of the jaw opening gesture for the last syllable of the two utterances (Figure 2, middle). Automatic labels were checked in order to correct errors and to detect trials for which one sensor was partially hidden in the optotrack. For the finger, the rule was to take the trajectory of the left sensor if the middle one was masked (the mean correlations between the two finger sensors were above .99 for each of the x-y-z coordinates).

Experimental design and hypothesis

The experiment manipulated two within-subject factors: the number of syllables (1 vs. 2 vs. 3 vs. 4) and the target position (near vs. far). As explained in the introduction, a (1=2<3=4) portrait should be observed for the period of pointing cycle. This period was computed for each trial as the duration between the apex of the two finger gestures. It will be referred as P_T:

$$P_T = P_{A2} - P_{A1}$$

Hence, the main hypothesis was that P_T should stay stable from the 1- to the 2- syllable(s) condition. It should then increase from the 2- to the 3- syllables condition and stay stable from the 3- to the 4- syllables condition. Two main questions would remain regarding the (1=2<3=4) portrait: (1) If actually displayed, does it resists to the increase of the pointing motion amplitude from the near- to the far- target position? (2) How is it achieved considering the different phases of pointing motion, that are the onset-to-apex duration and the post-apex phase? In order to give first element of answer to this question, the onset-to-apex pointing durations P_{D1} and P_{D2} were computed respectively for the first and second pointing gestures:

$$P_{D1} = P_{A1} - P_{I1} \qquad P_{D2} = P_{A2} - P_{I2}$$

These durations were compared with the total jaw motion duration necessary to produce the utterance, referred as J_{D1} and J_{D2} respectively for the first and second utterances:

$$J_{D1} = J_{A1} - J_{I1} \qquad J_{D2} = J_{A2} - J_{I2}$$

Results

Factor effects on dependent variables were tested using within-subject ANOVAs. As the number of syllables has more than two levels, sphericity was systematically tested (indicated only when it could not be assumed). Comparisons between 1- and 2- (C1), 2- and 3- (C2) and 3- and 4- (C3) syllables conditions were achieved using paired t-tests with Dunn-Sidák alpha-level adjustment in the case of a priori comparisons and with Bonferroni corrections in the case of post-hoc comparisons. Non-significant tests correspond to p-value greater than .05.

Apex-to-apex pointing duration (P_T)

Figure 3 displays P_T means and standard deviations for each experimental condition. Mauchly' sphericity test being significant for the number of syllables (Greenhouse-Geisser: Epsilon = .47. $p < .05$), ddl and p-values are given after Greenhouse-Geisser correction (which explains the non-integer values given for ddl). P_T mean is 961 ms for both 1- and 2- syllable(s) conditions while it increases to 1025 ms for the 3- and to 1056 ms for the 4- syllables conditions ($F(1.4, 11.2) = 9.3$, $p < .01$). Results of the three planned comparisons are: $t(8) = 0.05$, $p = .96$, for C1; $t(8) = 3.1$, $p = 0.015$, for C2 and $t(8) = 2.5$, $p = .04$, for C3. The Dunn-Sidák method shows that C3 is significant ($p < (1 - (1 - 0.05)^{1/3})$) but not C2 ($p > (1 – (1 – 0.05)^{1/2})$). This agrees with the (1=2<3=4) expected portrait. In addition, P_T is significantly longer in the far- (1022 ms) than in the near- (979 ms) target condition ($F(1, 8) = 13.9$, $p < .01$). Yet, the interaction with the number of syllables effect is not significant ($F(2, 15.8) = 1.7$). Hence, the +43 ms increase from the near to the far- target condition might be too small to affect the (1=2<3=4) clustering. The analysis will now investigate if this portrait is also observed for the onset-to-apex pointing duration.

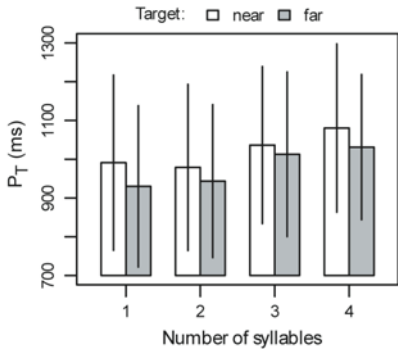

Figure 3: Means and standard deviations of the elapsed time between the apex of the two pointing gestures (P_T) according to the number of syllables and to the target position.

Onset-to-apex pointing duration (P_D)

Figure 4 displays P_D means and standard deviations for each experimental condition and respectively for the first (P_{D1}) and the second (P_{D2}) gesture. In the analysis, the gesture (first vs. second) was introduced as a third factor in a 3-within factors ANOVA (number of syllables * target position * gesture). The results show that durations for the first (364 ms) and the second (365 ms) gestures are very close to each other and do not significantly differ ($F(1, 8) = 0.02$). Then, gesture durations tend to be longer in the far- ($P_{D1} = 375$ ms, $P_{D2} = 379$ ms) than in the near- target condition ($P_{D1} = 352$ ms, $P_{D2} = 350$ ms). This target effect is significant ($F(1, 8) = 11.4$, $p < .01$) and does no significantly interact with gesture position ($F(1, 8) = 2.4$). Furthermore, P_D mean is about 356 ms for both the 1- and the 2- syllable(s) conditions, and increases to 369 ms in the 3- and to 376 ms in the 4-syllables conditions. The number-of-syllables effect is significant ($F(3, 24) = 4.4$, $p < 0.05$). However, C2 (+13 ms) and C3 (+7 ms) comparisons fail to reach significance. All other interactions are not significant.

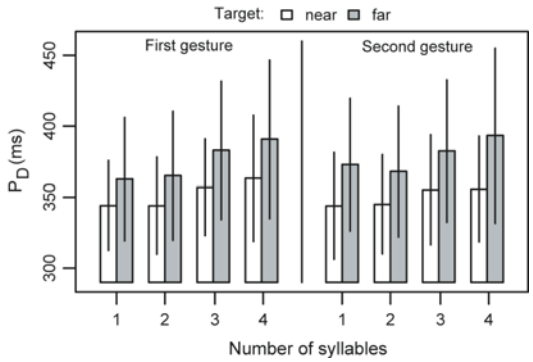

Figure 4: Means and standard deviations of the onset-to-apex pointing duration (P_D) according to the number of syllables and to the target position for the first (P_{D1}, left) and the second (P_{D2}, right) gestures

Comparison of jaw (J_D) and pointing (P_D) durations

Figure 5 displays the means of jaw motions durations (J_D) and standard deviations for each experimental condition and respectively for the first (J_{D1}) and the second (J_{D2}) utterance. As for P_D, the utterance (first vs. second) was introduced as a third factor for the ANOVA. The results show that jaw motion duration increases with the number of syllables. From the 1- to the 4- conditions, J_D is, respectively, 187, 390, 522 and 637 ms ($F(3, 24) = 566$, $p < .0001$). C1 (+203 ms), C2 (+132 ms), and C3 (+115 ms) are all significant ($(t8) > 16$, $p < .0001$). On the contrary, the target effect is not significant (near: 432 ms, far: 436 ms, $F(1, 8) = 3.6$).

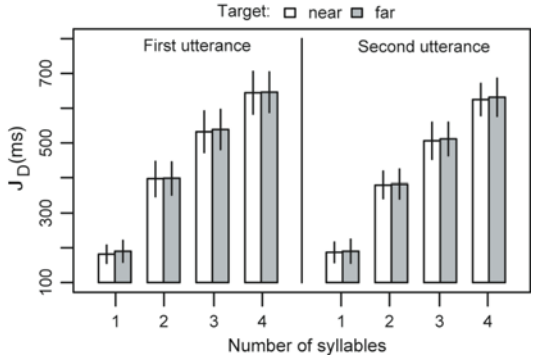

Figure 5: Means and standard deviations of jaw motion duration (J_D, see text for detail) according to the number of syllables and to the target position for the first (J_{D1}, left) and the second (J_{D2}, right) utterances.

The global comparison of J_D with P_D shows that the 365 ms mean observed for P_D is closer to the 390 ms mean observed for J_D in the 2- syllables condition than in the three other syllables conditions. The analysis of ($J_D – P_D$) shows that the mean of the ($J_D – P_D$) difference is -169, 34, 153 and 261 ms respectively from the 1- to the 4- syllable(s) conditions. These values significantly differ from zero at post-hoc t-tests (Bonferroni Correction) for the 1-, the 3- and the 4-

syllables condition ($t(8) > 9$, $p_{BF} < .0001$) but not for the 2-syllables one ($t(8) = 2.2$). Hence, the trend is that the duration of the pointing gesture corresponds rather closely with the duration of a sequence of two jaw cycles necessary for uttering a two-syllables component.

Discussion

Overall, the present findings agree with the 2:1 hypothesis. However, their interpretation has to be discussed in regard to a possible methodological problem: the eventuality of a higher processing load for /papapa/ and /papapapa/ than for /pa/ and /papa/.

Processing load

It could be suspected that the 1-2 vs. 3-4 clustering observed for P_T is due, at least partly, to a processing load higher for the motor programming of 3- vs. 4- syllables utterances. Indeed, the visual presentation of the items on the screen induced a trend to produce much confusion between these two kinds of sequences, while it was not the case for 1- and 2-syllable(s). In order to solve this problem, the experimenter gave a help to the subject by reading aloud the logatom, which avoided utterance errors. The impact of this intervention was reduced by the fact that the task is an "off-line" one: the subject waits for the go-signal to answer (e.g. Levelt et al. (1985)). However, this could have resulted in artificially clustering pointing durations at a high value, similar for the two kinds of sequences. If this was the case, the onset of pointing (P_{I1}) and jaw (J_{I1}) motions for the 3-4-syllables conditions might occur later than for the 1-2-syllable(s) conditions.

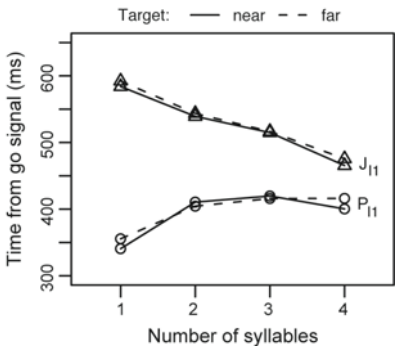

Figure 6: Means of elapsed time from the go signal to the pointing onset (P_{I1}) and to the onset of jaw opening gesture for the first vowel (J_{I1}) according to the number of syllables and to the target position.

Figure 6 displays the means of P_{I1} and J_{I1} according to the target position and to the number of syllables. The P_{I1} mean is 348, 408, 418 and 408 ms, respectively for the 1-, 2-, 3- and 4- syllables conditions. A 2-within-factors ANOVA shows that the effect of the number of syllables is significant ($F(3, 24) = 7.6$, $p < 0.001$) while the target effect is not significant ($F(1, 8) = 0.3$). Moreover C1 is significant ($t(8) = 3.1$, $p_{BF} < 0.05$) while neither C2 ($t(8) = 0.7$) nor C3 ($t(8) = 0.8$) are significant. Similarly, the number of syllables significantly affects J_{I1} ($F(3, 24) = 10.6$, $p < 0.001$) while it is not the case for the target position ($F(1, 8) = 1.4$). J_{I1} tends to decrease with the increase of the number of syllables: 588, 542, 516 and 471 ms, respectively for the 1-, 2-, 3- and 4-syllables conditions. However, C1 ($t(8) = 1.5$), C2 ($t(8) = 1.3$) and C3 ($t(8) = 2.8$) are not significant ($p_{BF} > .05$). Altogether, these patterns are not in favor of the "processing load" effect.

Arguments for the 2:1 hypothesis

The present study provides interesting results for the 2:1 hypothesis assumed in the Vocalize-to-Localize framework. Firstly, the oscillatory period of pointing motion is the same for 1- and 2- syllable(s) while it increases for 3- and 4-syllables. On the contrary, the onset-to-apex pointing duration is rather stable for a given target position. Yet, in agreement with the 2:1 ratio, the duration of this gesture corresponds rather well with the total duration of jaw motions for the realization of two CV syllables utterances. Hence, the (1=2<3=4) clustering observed for the whole pointing cycle may mainly result from the arm-hand-finger waiting for the jaw during the post-apex period in the 3- and 4- syllables conditions. In addition, because of the lack of gesture-alone condition in this preliminary experiment, we do not know if for the 1- and 2- syllables conditions, the arm-hand-finger system keeps its natural frequency. Nevertheless, previous data by Levelt et al. (1985) suggest that the duration of finger pointing motion with a 1- syllable utterance is close to the duration in a gesture-alone condition. Hence, two syllables might be the maximum number of syllables that could be realized on one finger pointing motion without affecting the duration of the pointing period. The lack of difference between the 3- and 4- syllables conditions and the proximity of values of jaw and pointing motion durations in the 2-syllables condition strengthen this assumption.

The 2:1 ratio as a new piece in language embodiment theories

Altogether, this preliminary experiment supports the assumption that the 2:1 ratio between speech and pointing observed in developmental studies (Ducey-Kaufmann et al., 2005) also tends to appear for adult productions. This adds new evidence for a close motor link between speech and hand gestures. Moreover, considering the importance of deixis in language acquisition and especially in first-words emergence and vocabulary expansion, it seems not so "astonishing" (Ducey-Kaufmann et al., 2005) to propose that this relationship between the durations of pointing and jaw gestures could have played a role in the course of phylogeny and may still play a role in the preference for bi-syllabic forms both in infants' first words and adults' lexicons (Rousset, 2004). Hence, we suggest that the 2:1 ratio assumption should be introduced as a new piece in computational models that attempt to derive phonology from substantial constraints (Lindblom, 1990, Steels, 2003, Schwartz et al., 2006). This integration might provide a basis for explaining the preference for bi-syllabic forms in human languages.

Conclusion

The present study leads to assume that two jaw cycles is the maximum number of cycles that could be realized inside a

pointing cycle without affecting the pointing duration. Moreover, it provides a new paradigm for the investigation of the coordination between the *Speech frame* and the *Sign frame*. Of course, more investigations are needed, particularly concerning the lack of pure vocal and gestural conditions, and further assessment of the processing load problem. This is why this study is presented as a preliminary investigation. In any case, further investigations about coordination between articulatory motion and hand gestures should provide new elements for a better understanding of the relationships between oral and sign communication, and of the real nature of human language.

Acknowledgments

This work is part of the "Patipapa" project funded by the French Ministry of Research (Action Concertée Incitative "Systèmes Complexes en Sciences Humaines et Sociales"). It benefited from inspiring discussions with Christian Abry.

References

Abry, C., Vilain A. and Schwartz J.L. (2004). Introduction: Vocalize to Localize? A call for better crosstalk between auditory and visual communication systems researchers: From meerkats to humans. In Abry C., Vilain, A. and Schwartz, J.L., editors, *Vocalize to Localize*, 313–325.

Arbib, M. A. (2005). Interweaving protosign and protospeech: Further developments beyond mirror. In C. Abry, A. Vilain & J.-L. Schwartz (Eds.) *Special issue: "Vocalize to Localize II". Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems, 6 (2)*, 145-171.

Corballis, M.C. (2003). From mouth to hand : Gesture, speech and the evolution of right-handedness. *Behavioral and Brain Sciences*, 26, 199-260.

Ducey-Kaufmann, V., Abry, C. & Vilain, C. (2005). When the Speech Frame meets the Sign Frame in a developmental framework. In *Proceedings of Emergence of Language Abilities: Ontogeny and phylogeny*, Lyon.

Emmory, K., Damasio, H., McCullogh, S., Grabowski, T., Ponto, L.L.B., Hichwa, R.D. and Bellugi, U. (2002). Neural Systems underlying spatial language in American Sign Language, *Neuroimage*, 17:812-24.

Feyereisen, P. (1997). The competition between gesture and speech production in dual-task paradigms. *Journal of Memory and Language*, 36(1):13-33.

Green, J.R., Moore, C.A., & Reilly, K.J. (2002). The sequential development of jaw and lip control for speech. *Journal of Speech, Language, and Hearing Research*, *45*, 66-79.

Haviland, JB. (2000). Pointing, gesture spaces, and mental maps. In McNeill, D., editor, *Language and gesture*, 13–46.

Iverson, J.M. and Goldin-Meadow, S. (1998), 'Why people gesture when they speak', *Nature*, 396, p.228.

Iverson, J.M. and Thelen, E. (1999). Hand, mouth, and brain: The dynamic emergence of speech and gesture. Journal of Consciousness Studies, 6:19-40.

Kelso, J., Tuller, B. & Harris, K. (1983). A "Dynamic Pattern" perspective on the control and coordination of movement. In: *The production of speech*, MacNeilage, P.F. (ed.), Springer Verlag: New York, pp. 137–173.

Levelt, W. J. M., Richardson, G. and La Heij, W. Pointing and voicing in deictic expressions. *Journal of Memory and Language*, 24:133-164, 1985.

Lindblom, B. (1990). *Explaining Phonetic Variation, A Sketch of the H H Theory*, pages 403-439. Academic Publishers.

MacNeilage, P.F. (1998).The frame/content theory of evolution of speech production. *Behavioral and Brain Sciences.*, 21:499-511.

MacNeilage, P.F. and Davis B.L. (2000). On the origins of internal structure of word forms. *Science*, 288:527-531.

McNeill, D. (1981). Action, thought and language. *Cognition* 10: 201-208

MacSweeney. M., Woll, B., Campbell, C., McGuire, P.K., Calvert, G.A., David, A.S., Williams, S.C.R. and Brammer, M.J. (2002). Neural systems underlying British Sign Language sentence comprehension. *Brain*, 125:1583-93.

Munhall K.G. and Jones J.A. (1998). Articulatory evidence for syllabic structure. *Behavioral and Brain Sciences*, 21:524-525.

Nelson, W.L., Perkell, J.L. and Westbury, J.R. (1984). Mandibule movements during increasingly rapid articulations of single syllables : Preliminary observations. *J. Acoust. Soc. Am.*, 75(3):945-951.

Nishitani N, Schurmann M, Amunts K, Hari R. (2005). Broca's region: from action to language. Physiology (Bethesda), Feb;20:60-9. Review.

Pizzuto, E., Capobianco, M., & Devescovi, A. (2005). Gestural-vocal deixis and representational skills in early language development. In C. Abry, A. Vilain & J.-L. Schwartz (Eds.) *Special issue: "Vocalize to Localize II". Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems, 6 (2)*, 223-252.

Pulvermuller. F. (2005). Brain mechanisms linking language and action. Natural Review of Neuroscience, 6(7):576-82.

Rousset, I. (2004). *Structures syllabiques et lexicales des langues du monde. Données typologiques, tendances universelles et contraintes substantielles*. Thèse en Sciences du Langage, Université Grenoble III.

San Jose-Robertson L, Corina DP, Ackerman D, Guillemin A, Braun AR. (2004) Neural systems for sign language production: Mechanisms supporting lexical selection, phonological encoding, and articulation, Human Brain Mapping 23(3), 156.

Schwartz, J.L., ., Boë, L.J., & Abry, C. (2006). Linking the Dispersion-Focalization Theory (DFT) and the Maximum Utilization of the Available Distinctive Features (MUAF) principle in a Perception-for-Action-Control Theory (PACT). In M.J. Solé, P. Beddor & M. Ohala (eds.) Experimental Approaches to Phonology. OUP (to appear).

Steels, L. (2003) Evolving grounded communication for robots. *Trends in Cognitive Sciences*, 7(7):308-312.

Treffner, P. and Peter, M. (2002). Intentional and attentional dynamics of speech-hand coordination. Human Movement Science, 21(5-6):641-97, 2002.

Volterra, V. , Caselli, M. C. , Capirci,O. , Pizzuto, E. (2005). Gesture and the emergence and development of language. In M. Tomasello and D. Slobin, (Eds.). *Beyond nature-nurture – Essays in honor of Elizabeth Bates*. Mahwah, N. J. : Lawrence Erlbaum Associates, pp. 3-40.