The Correlation of Speech and Hand Gestures for Multimodal Web Interaction

Jing Liu[1], **Manolya Kavakli**[2]
Department of Computing, Macquarie University, Sydney, NSW, Australia

Abstract - *With the development of Multimodal Interfaces (MMI) in Human Computer Interaction (HCI), there is an increasing interest at applying this technology to multimodal web interaction. Multimodal web interfaces can provide end-users with a natural, flexible and non-invasive interface that allow graphical, vocal and gestural interaction with web. Integration of speech and gestures in an MMI framework is now the focus of the researchers in this area. In order to combine speech and gestures in multimodal web interaction, it is essential to know the correlations between speech and associated gestures. This paper presents an empirical study aimed at studying the correlations between speech and hand gestures from a cognitive aspect. The methodology used in this paper is the video analysis to investigate the cognitive actions of speakers in the descriptions of objects using speech and hand gestures. The speakers' cognitive actions are analyzed using a cognitive scheme and protocol analysis method. Our initial findings suggest that speech is highly correlated with co-verbal hand gestures perceptually and semantically, regardless of the age, gender, background of the speakers, or the speed of speech and gesticulation.*

Keywords: multimodal web interaction, speech, co-verbal hand gestures, cognitive actions

1 Introduction

The Multimodal Interaction Activity is an initiative from W3C aiming to provide means to support multimodal interaction scenarios on the web. Multimodal interaction offers significant ease of use and benefits over uni-modal interaction from many aspects. Hands-free operation is needed in mobile devices with limited keyboards, as well as when a traditional desktop computer is unavailable to host the application user interface for controlling other devices. Multimodal web interaction is driven by the possibility in embedded and network-based speech processing for integrated multimodal web browsers. There is an exciting range of applications relevant to integrated multimodal web browsers. For instance, an ambient intelligent web interface is expected to add value to remote control of home entertainment systems. It can enable sensors, interactive screens, input devices for speech, gestures and tactile information to directly interact with each house and outdoor device. End-users with disability will also benefit from this technology [1][2]. VoiceXML is the W3C's standard XML format for specifying interactive voice dialogues between a human and a computer. It allows voice applications to be developed and deployed in an analogous way to HTML for visual applications [3]. With the emergence of MMI integrating speech and gestures [4][5], it is possible to combine these two input modes in web interfaces to get joint benefits. A system using personal robot as ubiquitous multimedia mobile web interfaces enable end-users to access web by speech and hand gestures [6]. The correlation of speech and hand gesture is crucial for integrating them in MMI as well as multimodal web browsers, since the architecture for MMI is time-sensitive for joint input.

2 Related work

A gesture is a form of non-verbal communication in which physical actions communicate particular messages, either in place of speech or together with speech. All speakers use gestures, although the typology of gesticulation may differ. They are tightly timed with speech [7]. Iconic gestures are found to precede the related speech within 2 seconds in [8]. Currently there are three different views about the relationship between speech and gestures. The first one points that speech and gestures are separately communicated [9]-[12]. According to this view, the primary role of gestures is to compensate for speech, when verbal communication is temporarily unavailable (e.g. coughing or hard to express by words). They argue that the process of gesture production has no effect on the process of speech production or the cognitive processes related to speech.

The second point of view is proposed initially by Robert Krauss [13][14]. It states that speech and gestures are linked reciprocally at a specific point during speech production. They point that the production of gestures is activated when speakers come across some difficulties in lexical retrieval. The activation of gestures in turn activates the lexical affiliate of that concept in mind, which results in articulating of the word successfully. According to this view, gestures are linked with speech only to the extent that it stimulates the activation of word retrieval in speech at a moment.

The third one articulated by David McNeill [7] argues that speech and gesture form an integrated system of communication. The links between speech and gesture are presented at the different levels of speech production (e.g. discourse, syntax, semantics and prosody). From this

Table I : Cognitive action categories

Category	Name	Description	Examples
Physical	D-action	Make depictions	Lines, circles, arrows, words
	L-action	Look at previous depictions	-
	M-action	Other physical actions	Move a pen, move elements, gesture
Perceptual	P-action	Attend to visual features of elements	Shapes, sizes, textures
		Attend to spatial relations among elements	Proximity, alignment, intersection
		Organize or compare elements	Grouping, similarity, contrast
Functional	F-action	Explore the issues of interactions between artifacts and people/nature	Functions, circulation of people, views, lighting conditions
		Consider psychological reactions of people	Fascination, motivation, cheerfulness
Conceptual	E-action	Make preferential and aesthetic evaluations	Like-dislike, good-bad, beautiful-ugly
	G-action	Set up goals	-
	K-action	Retrieve knowledge	-

standpoint, speech and gesture co-occur with one another during the same underlying thought process, even though the two modalities may capture and reflect different aspects of the common underlying cognitive process. The process of the productions of gesture and speech should therefore influence each other at any disrupted point.

There are actually already some neuropsychological and neurophysiologic evidence supporting the idea that speech and gesture share the same communicating system [15]. From the language aspect, some researchers hypothesize that language originate from an ancient system in which arm gestures is the communication tool [16]. Recently Corballis [17] propose that spoken language is developed as the repertoire of gestures gradually transferred from arm to mouth. From the gesture aspect, previous studies [7][12] show that pronouncing words and executing gestures with the same meaning are interacted and temporally coordinated. There is also evidence indicating that gesture has the impact on the utterance co-occurred. Kita also claims that gesture helps the speaker package information at an early stage of utterance production [18]. Conversely, gestures influence speech spectra of utterance produced simultaneously with the gestures [15].

In this paper, we are inspired by and expand the third view regarding the relationship between speech and gestures that they form a single communication system. Our hypothesis is that speech and co-occurring hand gestures are highly correlated to one another from cognitive aspect.

3 Types of gestures

According to [7], there are four main types of gesture regarding to their relationship to the concurrent speech. Deictic gestures mostly refer to actual entities and are used to specialize and locate in physical space. For example, imagine that you are communicating with a child and trying to tell him what the surroundings are. You normally say, 'Look, there is ... there' with a pointing gesture referring the object you mention. It may be hard for the child to recognize what you are talking about without the gesture. Iconic gestures mostly convey information about the outline of a picture of shape or object in space or the hands represent the shape or the object itself. These gestures are imagistically representational. Metaphoric gestures are also representational, but they are more associated with abstract ideas related to subjective notions, rather than the object itself. Beat gestures are small baton-like hand movements that serve to mark the speech pace normally. These gestures are not considered to convey any semantic information.

Among these four types, deictic gestures are probably the simplest gestures and beat gestures exhibit relatively little structural variation. Iconic gestures and metaphoric gestures are more complex than deictic and beat gestures with respect to both gesticulation and information they convey. Iconic gestures bear a close formal relationship to the semantic content of speech [19]. When we externalize imaginary environment of shapes and objects in our minds, it is a natural and intuitive way to use our hands as well as speech. Research has shown that the articulation of shapes and objects is performed using iconic gestures in both sign language and natural gestures [20]. This is also observed in our experiments. People used a variety of iconic gestures when they described the objects with tangible shapes.

4 Cognitive analysis

Cognitive analysis is the analysis of those properties of the objects that are accounted for in terms of cognitive concepts, such as various types of mental representation. It is to reveal the content what the speakers see, attend to, think of and retrieve from the memory by cognitive analysis of video/audio protocols. We used the content-oriented retrospective protocol analysis to investigate the cognitive actions of speakers. M. Suwa, T. Purcell, and J. Gero, [22] developed a coding scheme to code designers' cognitive actions. The scheme identifies various types of cognitive actions and reveals the structure of cognitive actions in designing process. We introduce the coding scheme briefly in the following part, before we illustrate how we used it in our experiments.

4.1 Coding scheme for cognitive actions

According to Suwa's coding scheme, cognitive actions of designers are classified into four information categories: physical, perceptual, functional and conceptual. Table I shows the detailed information about the four categories. M.

Table II : Codes of P-actions

Psg: discover a space as ground	Pfn: attend to the feature of a new depiction
Pfnp: attend to the feature of a new relation or Psg	Pfp: discover a new feature of an existing depiction, of Pcsg, or of Prsg
Prn: create or attend to a new relation between two new depictions or Psg	Prnp: create or attend to a new relation between a new depiction and an existing one
Prp: discover a spatial or organizational relation	Pcf: continually attend to a feature
Pcr: continually attend to a relation	Pcsg: continually attend to a space as ground
Prf: remember a feature of a depiction	Prr: remember a spatial or organizational relation
Prsg: remember a space as ground	Pipsr: implement a previously mentioned relation by giving new depictions or features

Table III : Codes of F-actions

Fnp: think of a function independently of depictions	Fre-i: re-interpretation
Fn: associate a new depiction, feature or relation with a new function	Fcp: continually think of a function independently of depictions
Fc: continually think of a function	Fr: remember a function
Frp: remember a function independently of depictions	Fi: implement a previously explored function by creating a new depiction, feature or relation

Suwa, T. Purcell, and J. Gero [22] claimed that these four categories are classified according to how it is processed by human cognition. Thus, physical actions correspond to sensory level at which incoming information is first processed sensorially. Then the incoming information is processed perceptually and semantically which are represented by perceptual actions, functional and conceptual actions respectively.

4.2 Codes of different actions

The coding scheme explored by Suwa etc. is based on the architectural designers' design activities. They detected a wide range of cognitive activities during the design session which is a complex task. The details about the procedures and coding can be found in [22][23]. In our experiments, what we are concerned is the correlation between speech and hand gestures. We therefore coded hand gestures as M-actions which represented by Mge. We expanded the Mge into four sub-classes: Mgei indicating iconic gestures, Mged corresponding to deictic gestures, Mgem for metaphoric gestures and also Mgeb for beat gestures. For the purpose of analysis, we will use G to represent gestures which include all M-actions. Perceptual and functional actions are also coded analyzing speakers' speech which is believed to reflect

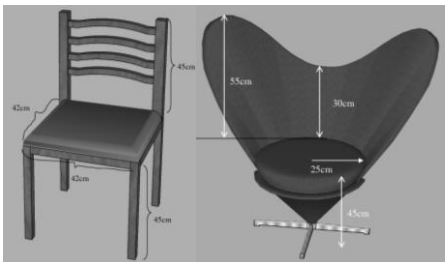

Figure 1. Classic chair (left) and Abstract chair (right)

cognitive thoughts in the speakers' mind. We rarely found conceptual actions in the protocols we captured. So, we only show the subcategories and codes of perceptual actions and functional actions in Table II and Table III. P-actions and F-actions are used to indicate perceptual actions and functional actions in the following.

5 Data and processiong

The hypothesis taken as a base for this study is that speech and co-verbal hand gestures are correlated with each other at the perceptual level as well as the semantic level. The correlation has nothing to do with the age, the gender and the background of the speakers, nor the speed of the speech and gesticulation. In order to explore the correlation between speech and co-verbal hand gestures we conducted our own experiments. In our experiments, the participants were required to describe two chairs (See Fig.1) with different structures, as if they are having a video conference with someone. Protocol analyses are carried out on the videos/audios we captured.

5.1 Corpus and data

This study is based on the analysis of a set of multimodal corpus collected by ourselves. 16 volunteers (aging from twenties to fifties, including 12 females and 4 males) were involved in the data collection for the experiment. They are from different cultural backgrounds but speak English in our experiment. The participants were required to describe two different types of chairs naturally to the camera. One is a traditional chair with a simple structure. Another one is designed with an abstract shape and structure, which is expected to trigger more gestures of. We gave 3D pictures about both chairs to the participants. They described the objects freely in front of a camera. The camera was placed in such a way that the hands and the upper body gestures could be recorded clearly. All their speech was recorded with the video camera using internal microphone. Finally, we obtained approximately 30 minutes of monologue object description data.

5.2 Coding process

Participants in the experiment were encouraged to use as many gestures as possible. The analysis is via two annotations

tool: Anvil [24] for video annotation and Praat[25] for audio annotation.

5.2.1 Segmention

Segmentation is the first step of the coding process. There are different rules for segmentation. One rule is that protocols can be segmented by verbalization events (e.g. pauses, intonations and syntactic markers). But for cognitive segments, we divided the protocols based on cognitive actions which reflect the subject's intention. We identified a new segment when there is a change in the speakers' intention or the contents of their thoughts. For example, a participant may have said, "the seat of this chair is square… and then for the leg part ..."

The speaker changes his/her attention from the seat part to the leg part. That case we define the start point of a new segment as 'and then ...' consequently. A single segment can include one sentence or many.

5.2.2 Gesture coding

We first analyzed the videos and segmented the video footage while listening to the speech to ensure that we obtained the starting and ending frame of the segmentation. For gesture analysis of each segment, we recorded the gesture types (iconic, metaphoric, beat, deictic) corresponding to McNeill's classification.

We extracted gestures from the video protocols based on the ANVIL built-in gesture phase descriptions. In the description file, a gesture is divided into 7 phases: (Pre, Stroke, Hold, Beats, Recoil, Partial-retract, Retract)(quoted from help pages in ANVIL). In our analysis, gestures are represented only by 'stroke' phase, because the stroke phase is the most energetic part of a gesture movement and also the requisite part of a gesture. The movement for a gesture stroke was often apparent in the video frames as a blurring of the hands; the cessation of the blurring in one stroke movement was taken as the end of a gesture [26]. Other phases were not recorded since the beginnings of other phases for each gesture were subject to greater subjectivity and difficulty in identification.

5.2.3 Cognitive action coding

Coding of cognitive actions was finalized by speech analysis via the annotation tool Praat. By Praat, we were able to analyze the speech with the display of speech intensity contour. This is helpful in pinpointing the start and end point of a segment as well as the words related to the gestures. Praat also allows users to rehear any selected part of the audio (e.g. one segmentation) unlimited times to make the coding more reliable.

We illustrate a coding example for one segment as follows.

The following sentences were excerpted from one participant's description about the traditional chair: "*Then the back of the chair very much straights up from the back leg. There are four strips going across and each stripe is curved a little bit and arched like this*"

Table IV. Coding example for one segment

M-actions		F-actions		P-actions	
M_{gei}	straight up	F_n	back	P_{fp1}	straight up
M_{gei}	strips	F_c1	numerical info 4	P_{fp2}	square
M_{geb}	4	F_c2	stripes	P_{fp3}	curved
M_{gei}	curved			P_{cf}	arched
				P_{rp}	going across

The excerpted part was about the back of the chair. Before this segment the participant mentioned the legs and after it he talked about the seat of the chair.

Four gestures were detected while we coded this segment. The participant gestured when he said 'straight up', 'four, 'strips' and 'curved'.

According to the coding scheme described above, we coded the cognitive actions for this segment as shown in Table IV.

6 Results

Approximately 30 minutes of monologue object descriptions in our video footage were obtained for the total 16 participants. We used seconds as a time measure. A total of 1974.02 seconds (from 42.73 seconds to 365 seconds for different participants) of video footage captured for analysis. In total, 100 segments (from 2 segments to 13 segments for different participants) were coded after segmentation.

We examined the frequency with which gestures, perceptual actions and functional actions occurred throughout the object description process of speakers. For each participant, we calculated the total of occurrences of physical, perceptual and functional actions. Table V displays the occurrences of these three actions for different speakers. From the table we can see, we obtained 13 segments for the participant (P1) and only got 2 segments for P16. P1 gestured 55 times during the whole process while P13 only produced 8 gestures. We detected the maximum numbers of P-actions and F-actions for P1 (49 and 44) respectively and the minimum for P16 (only 5 and 7 respectively). The reason for the difference could be the degree of comfort in using English to talk or describe complexity. We had some participants who come from non-English speaking countries but were required to speak in English.

It can be seen from the table that in our experiments, P2 had 24 gestures produced which is less than P8's (32), but we detected 24 P-actions for her which is more than P8's (22). 17 F-actions were coded for P12 which is more than P5's (15), but we obtained 17 P-actions and 15 gestures for her which are both less than P5's (21 and 25). However, we still can observe that the whole trend of these three actions is quite close to one another. This can be clearly seen in Fig 2. From this figure we can see the lines representing gestures, P-actions and F-actions are very close to each other, even though they cross lines at some points.

We calculated the correlation coefficients between the number of segments and each type of actions throughout all

Table V. Numbers of actions for participants

P	Sum(Seg)	Sum(G)	Sum(P)	Sum(F)
P1	13	55	49	44
P2	9	24	24	15
P3	13	43	41	25
P4	4	19	16	10
P5	6	25	21	15
P6	6	16	17	17
P7	4	11	11	8
P8	7	32	22	29
P9	6	21	22	26
P10	8	18	21	17
P11	4	10	11	11
P12	5	15	17	17
P13	3	8	7	10
P14	6	16	12	17
P15	4	15	8	11
P16	2	11	5	7

participants. As can been seen from Table.VI, there are strong correlations between gestures and different types of cognitive actions. Gestures are strongly correlated with perceptual actions (0.9566) as well as functional actions (0.8834). In addition to this, the segments are also strongly correlated with gestures (0.9028) and perceptual actions (0.9549). However, we found that segments are not so strongly correlated with functional actions (0.7988).

As we introduced before, for the same task, different speakers finished the experiments at different times (shortest time: 42.73 seconds and longest time: 365 seconds). As mentioned before, the participants were from different backgrounds (e.g. some are Australian local speakers and some are Asian speakers). The age of these participants vary from twenties to fifties. The speed of their talking and gesticulation are also varying. However, the correlation coefficients are calculated without calculating the effects of these factors. We can now state that gestures are strongly correlated with perceptual actions and functional actions without considering any individual differences.

Now we may relate gestures to speech from the cognitive aspect. We believe that speech is the important and reliable reflection of speakers' cognitive thoughts. The definition of the coding scheme for cognitive actions implied that the perceptual processing of incoming information is indicated by perceptual actions. Functional actions can reflect the semantic processing of incoming information. Therefore, the information we extracted from speech can reflect speakers' cognitive actions at different levels. After analyzing the codes for gestures and cognitive actions, we found that speech and co-occurring hand gestures are correlated with one another perceptually as well as semantically. This relationship is neither affected by the age, the gender, the background of the speakers nor the speed of the speech and gesticulation.

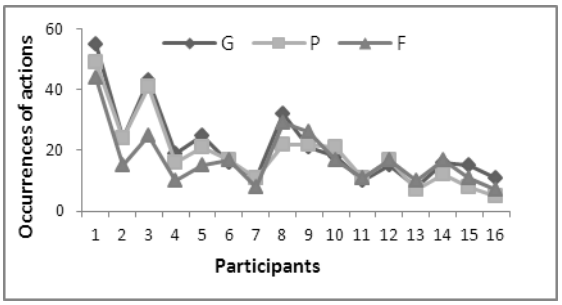

Figure 2. Occurrences of actions for participants

Table VI. Correlation coefficients

	G	P	F	Seg
G	1.000	0.9566	0.8834	0.9028
P	0.9566	1.000	0.8651	0.9549
F	0.8834	0.8651	1.000	0.7988
Seg	0.9028	0.9549	0.7988	1.000

7 Conclusion and future work

7.1 Conclusion

We started this paper in order to explore the correlation between speech and co-verbal hand gestures from the cognitive aspect for multimodal web interaction. The view of that speech and co-occurring hand gestures share the same communication system is not new. But a few researchers studied their relationship from the cognitive aspects by statistical analysis of cognitive. By making use of a coding scheme for designers' cognitive actions, we examined videos/audios of speakers which recorded speech and gesture information of speakers. We analyzed cognitive actions of speakers using the coding scheme and investigated the number of gestures and related P-actions and F-actions of various participants in our experiments. Our conclusion is that speech and hand gestures are strongly correlated, from the perceptual aspect as well as the semantic aspect. This was visible with the correlation coefficients of gestures and cognitive actions in participants (0.9566 for gestures and P-actions and 0.8834 for gestures and F-actions). There are already some researches which enable end-users to access web by speech. Considering that speech and hand gestures are highly correlated, we can expect a multimodal web which has web pages one can speak to and gesture at in future.

7.2 Limitations and future directions

We acknowledge that future work needs more samples. Future studies may need to work on the consistency of the annotations and codes. This was not possible to perform for this paper. Only one person was working on the annotation of gestures and the coding of cognitive actions. Further work can also address how to fuse gesture into multimodal web interaction.

8 Acknowledgement

This work was supported in part by the Australian Research Council Discovery grant DP0988088 to Manolya Kavakli, titled "A Gesture-Based Interface for Designing in Virtual Reality"

References

[1] González, Julia, Mercedes Macías, Roberto Rodríguez, and Fernando Sánchez. "Accessibility metrics of web pages for blind end-users." In Web Engineering, pp. 374-383. Springer Berlin Heidelberg, 2003.

[2] Knudsen, Lars Emil, and Harald Holone. "A multimodal approach to accessible web content on smartphones." In Computers Helping People with Special Needs, pp. 1-8. Springer Berlin Heidelberg, 2012.

[3] Larson, James A. "W3c speech interface languages: Voicexml [standards in a nutshell]." *Signal Processing Magazine, IEEE* 24.3 (2007): 126-131.

[4] Larson, James A. "W3c speech interface languages: Voicexml [standards in a nutshell]." *Signal Processing Magazine, IEEE* 24.3 (2007): 126-131.

[5] Wachs, Juan Pablo, Mathias Kölsch, Helman Stern, and Yael Edan." Vision-based hand-gesture applications." *Communications of the ACM* 54, no. 2 (2011): 60-71.

[6] Ruiz-del-Solar, Javier. "Personal robots as ubiquitous-multimedial-mobile web interfaces." In Web Conference, 2007. LA-WEB 2007. Latin American, pp. 120-127. IEEE, 2007.

[7] D. McNeill, Hand and Mind: What Gestures Reveal about Thought. Univ. Chicago Press, 1992.

[8] Liu, J. and Kavakli, M. "Temporal Relation between Speech and Co-verbal Iconic Gestures in Multimodal Interface Design", Retrieved April, 2012

[9] B. Butterworth and G. Beattie, "Gestures and silence as indicators of planning in speech," in Recent Advances in the Psychology of Language: Formal and Experimental Approaches, ed., 1978

[10] U. Hadar, "Two types of gesture and their role in speech production," Journal of Language and Social Psychology, vol. 8, no. 3-4, pp. 221–228, 1989.

[11] U. Hadar, D. Wenkert-Olenik, R. Krauss, and N. Soroker,"Gesture and the processing of speech: Neuropsychological evidence," Brain and language, vol. 62, no. 1, pp. 107–126,

[12] W. Levelt, G. Richardson, and W. La Heij, "Pointing and voicing in deictic expressions," Journal of Memory an Language, vol. 24, no. 2, pp. 133–164, 1985.

[13] R. Krauss, "Why do we gesture when we speak?" Current Directions in Psychological Science, vol. 7, no. 2, pp. 54–60, 1998.

[14] R. Krauss and U. Hadar, "The role of speech-related arm/hand gestures in word retrieval," Gesture, speech, and sign, pp. 93–116, 1999..

[15] P. Bernardis and M. Gentilucci, "Speech and gesture share the same communication system," Neuropsychologia, vol. 44, no. 2, pp. 178–190, 2006.

[16] D. Armstrong, W. Stokoe, and S. Wilcox, Gesture and the Nature of Language. Cambridge University Press, 1995.

[17] M. Corballis, From hand to mouth: The origins of language. Princeton University Press, 2003.

[18] S. Kita, "How representational gestures help speaking," Language and gesture, pp. 162–185, 2000.

[19] A. Erol, G. Bebis, M. Nicolescu, R. Boyle, and X. Twombly, Vision-based hand pose estimation: A review," Computer Vision and Image Understanding, vol. 108, no. 1, pp. 52–73, 2007.

[20] T. Marsh, "An iconic gesture is worth more than a thousand words," in Information Visualization, 1998. Proceedings. 1998 IEEE Conference on. IEEE, 1998, pp. 222–223.

[21] T. A. van Dijk. (2000) Cognitive discourse analysis. http://www.discourses.org/UnpublishedArticles/cogn-dis-anal.htm. University of Amsterdam.

[22] M. Suwa, T. Purcell, and J. Gero, "Macroscopic analysis of design processes based on a scheme for coding designers' cognitive actions," Design studies, vol. 19, no. 4, pp. 455–483, 1998.

[23] M. Suwa, J. Gero, and T. Purcell, "Analysis of cognitive processes of a designer as the foundation for support tools," in Artificial Intelligence in Design, vol. 98, 1998, pp. 229–248.

[24] M. Kipp. (2000-2012) The video annotation research tool. http://www.anvil-software.org/.

[25] P. Boersma and D. Weenink. The video annotation research tool. University of Amsterdam.

[26] Y. Yasinnik, M. Renwick, and S. Shattuck-Hufnagel, "The timing of speech-accompanying gestures with respect to prosody," Proceedings of Sound to Sense, MIT, 2004.