Gesture Patterns during Speech Repairs

Lei Chen, Mary Harper
Speech processing Lab
Electrical and Computer Engineering
Purdue University
West Lafayette, IN 47906-1285
{chenl, harper}@ecn.purdue.edu

Francis Quek
VisLab
Computer Science and Engineering
Wright State University
Dayton, OH 45435
quek@cs.wright.edu

Abstract

Speech and gesture are two primary modes used in natural human communication; hence, they are important inputs for a multimodal interface to process. One of the challenges for multimodal interfaces is to accurately recognize the words in spontaneous speech. This is partly due to the presence of speech repairs, which seriously degrade the accuracy of current speech recognition systems. Based on the assumption that speech and gesture arise from same thought process, we would expect to find patterns of gesture that co-occur with speech repairs that can be exploited by a multimodal processing system to more effectively process spontaneous speech.

To evaluate this hypothesis, we have conducted a measurement study of gesture and speech repair data extracted from videotapes of natural dialogs. Although we have found that gestures do not always co-occur with speech repairs, we observed that modification gesture patterns have a high correlation with content replacement speech repairs, but rarely occur with content repetitions. These results suggest that gesture patterns can help us to classify different types of speech repairs in order to correct them more accurately [6].

1. Introduction

Natural human communication is inherently multimodal. Two primary modes used in natural human communication are speech and gesture [4, 2]. As such, they are ideal inputs for multimodal human-to-computer interfaces to process. Computer understanding of natural human dialog is currently an unsolved problem, albeit an important one. One reason for this is that accurate machine understanding of spontaneous speech is still a highly challenging open problem for speech researchers. Understanding of spontaneous speech is a difficult problem because it often contains mistakes that are revised or repaired within the same utterance. The presence of these speech repairs contributes to low speech recognition accuracy of spontaneous speech in today's speech recognition systems [5, 15]. Speech repairs are often broken down into three components: the repair site or reparandum, the editing phrase (i.e., a spoken cue phrase like *"I mean"*), and the resumption site or alteration [6, 7, 12, 13]. Explicit editing phrases are not always present, and these phrases do not uniquely signal the presence of a repair [6]. Speech repairs are often classified [6, 7] as follows:

1. **false starts**: (also called fresh starts), e.g., the following example contains an utterance that is aborted by the speaker:

   [*I need to send*]$_{\text{aborted utterance}}$ [*let's see*]$_{\text{editing phrase}}$ [*how many boxcars can one engine take*]$_{\text{new utterance}}$.

2. **content replacements**: (or modification repairs), for example, the content of the reparandum is replaced by the alteration in the following example:

   you can [*carry them both on*]$_{\text{reparandum}}$ [*tow them both on*]$_{\text{alteration}}$ *the same engine*.

3. **repetitions**: e.g., [*she*]$_{\text{reparandum}}$ [*she*]$_{\text{alteration}}$ *liked it.*

Based on whether or not the content has been modified, we classify speech repairs as *content modifications* (i.e., false starts and content replacements) and *content repetitions* (i.e., repetitions).

A second reason that computer understanding of human-to-human dialog is difficult is that computer models for processing gestures and decoding their meaning are only beginning to surface. If we are to build systems that are able to exploit gesture in natural interaction, it is essential to derive computationally accessible metrics that can be utilized for processing the communication. Human communication is a dynamic interplay among various `communicative channels' that include speech, prosody, gesture, gaze, facial expression and body posture [10]. These modalities do not function independently, nor is any modality subservient to another. Instead these modalities proceed from the same thought process that produces an utterance, and each carries aspects of the original thought [11]. Based on the assumption that speech and gesture proceed from the same thought

process, we would expect to find patterns of gesture that co-occur with speech repairs.

2. Experimental Method

Data used in this paper came from KDI visual-audio database. Videos were recorded in a series of elicitation experiments performed at Department of Psychology at University of Chicago under the direction of David McNeill. Subjects were recruited in speaker-interlocutor pairs. To avoid `stranger-experimentor' inhibition in the discourse, the subjects already knew one another. The subject was shown a model of a village and told that a family of intelligent wombats have taken over the town theater, and was made privy to a plan to surround and capture the wombats and return them to Australia. This plan involves collaboration with the villagers, paths of approach, and encircling strategies. The subject was then videotaped communicating these with his/her interlocutor using the town model. The task description we use for the experiment is shown in Figure 1.

A family of intelligent wombats has taken up residence in an abandoned movie theater in the town of Atlee. You and your assistant need to catch the wombats so that you can send them back to Australia. You will be taking the train to Atlee, so be ready to get off at the station right after you pass a church on your right. When you get off the train, go around the station and cut through the adjacent park to meet your assistants: pass between the two trees and you should reach the house number 33. Then go next door and ask the neighbors in 35 (you'll notice the road construction in front of the house) to assist you. The movie theater is across the intersecting street. One of you should go in the front entrance and scare the wombats out the back entrances. With the help of the people in 33 and 35, you should be able to snare the wombats as they exit the rear entrance. Explain the task to your assistant and decide on who does what, and what equipment you will need to bring with you.

Figure 1. Task description used in our experiment.

We apply a three camera setup in our experiments. Two of the cameras are calibrated so that once correspondence between points in the two cameras is established, the 3D positions and velocities can be obtained. The third camera is a closeup of the head. We chose this configuration because our experiment must be portable and easy to set up (some of our cross-disciplinary collaborators collect data in the field). Using a five foot-wide prism with a known constellation of points, we are able to obtain points with typical average errors within 1 mm in x and y and about 1.5 mm in z (toward the cameras). The maximal errors are within 4 mm. We believe that this is sufficient for conversational gesture interaction. We use off-the-shelf consumer-grade miniDV 30 frames-per-second cameras in progressive scan mode in these experiments. The audio for each participant was digitally recorded using a Shure Sm94 unidirectional boom mounted microphone that was placed at a distance of eight inches from the subjects' mouths. The video and audio are synchronized using a movie `clapper' device. The video is digitalized on the SGI workstation and saved as SGI MPEG format. The audio was initially sampled at 44.1K and then downsampled to 14,700 KHz for analysis. One frame of video is shown in Figure 2.

Figure 2. One frame of video in KDI visual-audio database.

All videos are processed in the VisLab at Wright State University. A fuzzy image processing approach known as *Vector Coherence Mapping* VCM is used to track the hand motion [14]. VCM applies spatial coherence, momentum (temporal coherence), speed limit, and skin color constraints in the vector field computation by using a fuzzy-combination strategies, and produces good results for hand gesture tracking. An iterative clustering algorithm is applied that minimizes spatial and temporal vector variance to extract moving hands. The positions of the hands in the stereo images are used to produce 3D motion traces describing the gestures. Three gesture features are extracted for each hand of a speaker:

1. 3D Hand Position: the (x, y, z) position of a hand.

2. Hold: a state when there is no hand motion beyond some threshold. A motion energy-based detector was used to locate places where there was low motion energy [3].

3. Effort: analogous to the kinetic energy of hand movement [3].

We also performed a text transcription of each discourse that was initially time-aligned to the audio data using *Entropic Aligner* [16] and then modified by an experienced speech scientist using the Praat tool [1]. After the words were aligned, additional annotations of the data were added as Praat tiers, e.g., utterances, speech repairs, and discourse structure. Figure 3 depicts the annotation of a speech repair.

For this experiment, we selected two data sets from the KDI visual-audio database, wd11 and wd20, for analysis. Table 1 indicates the length of each data set in seconds (s) and provides information on each conversant, i.e., gender, the number of utterances produced (# Utt.), and the number of speech repairs produced (# Rep.).

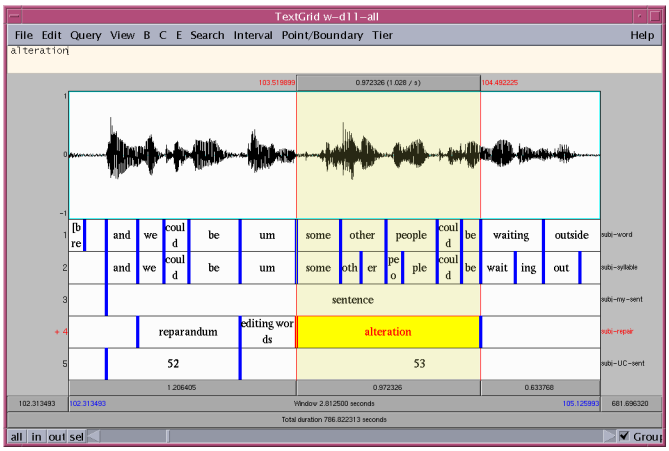

Figure 3. The annotation of a speech repair from the wd11 data set of the KDI visual-audio database.

Data	Length	Speaker	Gender	# Utt.	# Rep.
wd11	786.82 s	spk1	Female	277	44
		spk2	Male	220	16
wd20	727.20 s	spk1	Male	295	31
		spk2	Male	306	27

Table 1. A summary of the properties of the wd11 and wd20 data sets.

Data	Speaker	# Rep.	# Gesture-on	# PM
wd11	spk1	44	15	8
	spk2	16	4	0
wd20	spk1	31	9	4
	spk2	27	8	1

Table 2. A summary of the co-occurrence of gestures and speech repairs in wd11 and wd20.

3. Data Analysis and Results

For this investigation, we evaluate the types of gestures that co-occur with speech repairs by analyzing the transcription and corresponding video frames of each speech repair. We first determine whether a gesture co-occurs with each speech repair; if it does then the speech repair is in the **gesture-on** group. Note that in this initial investigation we do not consider movements such as touching glasses or hair or posture adjustment to be gestures; we call these pragmatic movements. Also, speech repairs that contain a very long pause between the reparandum and alteration are not considered[1].

Table 2 provides information on the co-occurrence of gesture and speech repairs in the wd11 and wd20 data sets. The number of times that a gesture co-occurs with a speech repair appears in the column labeled **# Gesture-on**. We also indicate the number of times a pragmatic movement (PM) co-occurs with a speech repair in the column labeled **# PM**.

Clearly, speakers do not always use gestures during speech repairs. An important question is, under what circumstances do they utilize gesture during a repair? In order to get a deeper understanding of gestures that occur during speech repairs, we re-analyzed the data to identify the different types of gestural patterns that occur during speech repairs in our data sets. To get a better understanding of the process we used for the analysis, consider the speech repair appearing in Figure 3 with the following transcription:

$$[\textit{we could be}]_{\text{reparandum}}\ [\textit{um}]_{\text{editing phrase}}\ [\textit{some other people could be}\ ]_{\text{alteration}}$$

Component	Words	Begin	End
reparandum	we could be	102.68	103.22
editing phrase	um	103.22	103.52
alteration	some other	103.52	104.50
	people could be		

Table 3. The begin and end time of each component of our example speech repair.

Information on the start and finish time of the reparandum, editing phrase, and alteration of our example repair appears in Table 3. To determine the types of gesture patterns occurring during the reparandum, editing phrase, and alteration of the speech repair, we stepped through the video on a frame-by-frame basis. Figure 4 shows some of the important key frames for our example speech repair. The gesture features

[1] There was only one instance in this study.

Figure 4. Key frames of gestures during our speech repair example.

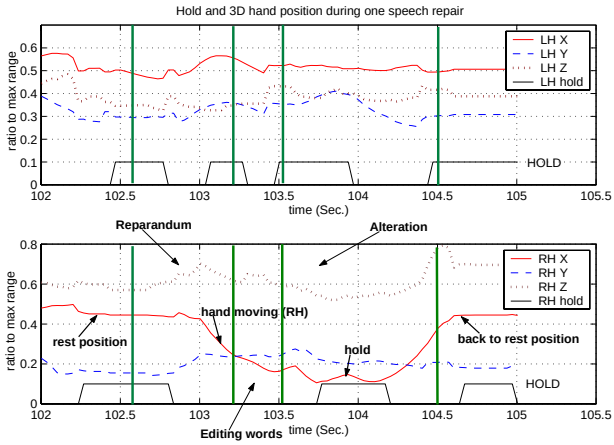

Figure 5. Left hand (LH) and right hand (RH) 3D positions and holds during the video sequence of our speech repair example.

for our example obtained by VCM are shown in Figure 5. The figure depicts at each time point whether the left (right) hand is moving or at a hold along with its X, Y, and Z coordinate. We also mark the start and finish times of the reparandum, editing phrase, and alteration. At the beginning of the reparandum, the left and right hand are both at rest on speaker's knees, as can be seen in the first key frame of Figure 4 and the VCM features of Figure 5 (e.g., notice the right hand's X position (RH X)). After a brief time, the speaker raises her hand and moves it away from her body. This movement continues through the editing phrase, and then in the alteration, both hands stop and then are retracted back to resting position.

Our example speech repair has a speech repair gesture pattern that we have observed during other speech repairs. Over all of our data, six distinct speech repair gesture patterns were observed:

1. **Case 0**: No gesture occurs during the reparandum, editing phrase, and alteration of the speech repair.

2. **Case 1**: No gesture appears in the reparandum, but a gesture begins in the alteration.

3. **Case 2**: A gesture is terminated within the reparandum and a new gesture begins in the alteration.

4. **Case 3**: A single gesture appears across the reparandum and alteration with a very slight hesitation of hand movement in the editing phrase.

5. **Case 4**: A gesture occurs in the reparandum, followed by a hold after the interruption point and a retraction of hands to rest in the alteration.

6. **Case 5**: A single gesture appears across the reparandum and alteration.

Our example speech repair is clearly Case 4. Table 6 provides an analysis of all of the speech repair gesture patterns that occur for spk1 of the wd11 data set.

Speech repairs reflect the need for some modification in the speech production process. The case 1, 2, and 4 gesture patterns exhibit a change in gesture state during either reparandum, alteration, or both; hence, we call them modification gestures. The simultaneous production of modification gestures in the visual channel along with the speech repair in the audio channel suggest that these channels are highly correlated. To understand why not all speech repairs exhibit modification gesture patterns (MG), we analyzed the distribution of modification gestures for the two major classifications of speech repairs, namely content modification (CM) and content repetition (CR) repairs.

The distribution of gestures in two kinds of speech repairs appear in Table 4. We find that gesture entities have high correlation with content modification speech repairs.

Data	Speaker	# CM	# MG	# CR	# MG
wd11	spk1	24	13	20	0
	spk2	7	4	9	0
wd20	spk1	12	8	19	1
	spk2	15	8	12	0

Table 4. The distribution of modification gestures (MG) with CM (content modification) and CR (content repetition) speech repairs.

Looking at the wd11-spk1 data set, we further refined the analysis of the content modification repairs into false starts (a total modification) and content replacements (a partial modification). The results are shown in Table 5. It is quite interesting that both types of speech repairs have a simi-

lar distribution of co-occurrence with a modification gesture pattern.

Some psychologists hypothesize that gestures help speakers to formulate coherent speech by aiding in the retrieval of elusive words from lexical memory [8]. Based on this hypothesis, when speech repairs are content replacements, one would expect a greater number of modification gestures, which could help in retrieving words. On the other hand, when speech repairs are simple word repetitions, it is more likely that the speaker is maintaining the conversational floor; hence, fewer gestures would be expected in this situation. Levelt [9] distinguishes between three stages of speech production, i.e., *conceptualization*,*formulation*, and *articulation*. During conceptualization, the speaker determines the communicative intention to produce the *preverbal message*. At the formulation stage, the preverbal message is transformed into a *surface structure*. Finally, speech is generated during the articulation stage from the surface structure [8]. Any disruption of the process is likely to incur a high retrieval cost; hence, we would expect that modification gestures would be used frequently for false starts and modification repairs, which signal such a disruption in processing.

4. Conclusion

In this paper, we have reported the results of a measurement study of gesture and speech repair data extracted from videotapes of natural dialogs. From our analysis of this data, we found that gestures do not always co-occur with speech repairs. However, we observed a set of six gesture patterns that occur during speech repairs, three of which we have dubbed modification gesture patterns due to the fact that there is a change of gesture state in either the reparandum, alteration, or both. Using this class of gesture patterns, we found that content modification speech repairs have a high correlation with modification gesture patterns, unlike content repetitions, which have very few co-occurring modification gestures. In future research work, we will examine more data sets and then begin to build speech repair detection algorithms that utilize gesture pattern as an additional knowledge source.

Repair	Content Replacement	False Start
	20	4
# MG	10	3

Table 5. The distribution of modification gestures occurring with content replacements and false starts.

5. Acknowledgements

This research was supported by Purdue Research Foundation and the National Science Foundation under Grant No. #9980054-BCS. Additional thanks go to our KDI grant colleagues.

References

[1] P. Boersma and D. Weeninck. Praat, a system for doing phonetics by computer. Technical Report 132, University of Amsterdam, Inst. of Phonetic Sc., 1996.

[2] R. A. Bolt. Put that there: Voice and gesture at the graphics interface. *ACM Computer Graphics*, 14:262–270, 1980.

[3] R. Bryll, F. Quek, and A. Esposito. Automatic Hand Hold Detection in Natural Conversation. In *IEEE Workshop on Cues in Communication, Kauai, Hawaii*, 2001.

[4] J. Cassell. A framework for Gesture Generation And Interpretation. In R. Clipolla and A. Pentland, editors, *Computer vision in Human-Machine Interaction*. Cambridge University Press, 1998.

[5] J. J. Godfrey, E. C. Holliman, and J. McDaniel. SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of the International Conference of Acoustics, Speech, and Signal Processing*, pages 517–520, 1992.

[6] P. A. Heeman and J. F. Allen. Speech repair, intonational phrases, and discourse markers: Modeling speakers' utterances in spoken dialogue. *Computational Linguistics*, 25(4):525–571, 1999.

[7] D. Hindle. Deterministic parsing of syntactic non-fluencies. In *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics*, pages 123–128, 1983.

[8] R. M. Krauss. Why do we gesture when we speak. *Current Direction in Psychological Science*, 7:54–59, 1998.

[9] W. J. Levelt. *Speaking: From intention to articulation*. The MIT Press, Cambrige, MA, 1989.

[10] D. McNeill. *Hand and Mind:What Gestures Reveal about thought*. U.Chicago Press, Chicago, IL, 1992.

[11] D. McNeill and S. Duncan. Growth Point in thinking-for-speaking. In D. McNeill, editor, *Language and Gesture*, pages 141–161. Cambridge University Press, 2001.

[12] C. H. Nakatani and J. Hirschberg. A speech-first model for repair detection and correction. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 46–53, 1993.

[13] C. H. Nakatani and J. Hirschberg. A corpus-based study of repair cues in spontaneous speech. *Journal of the Acoustical Society of America*, 95(3):1603–1616, 1994.

[14] F. Quek, R. Bryll, and X.-F. Ma. A Parallel Algorighm for Dynamic Gesture Tracking. In *ICCV'99 Workshop on RATFG-RTS, Corfu, Greece*, 1999.

[15] E. Shriberg and A. Stolke. Word probability after hesitations: A corpus-based study. In *Proceedings of the International Conference on Spoken Language Processing*, volume 3, pages 1868–1871, 1996.

[16] C. Wightman and D. Talkin. *The Aligner*. Entropic Research Laboratory,INC.

Number of Repairs	Repair Type	Reparandum	Editing Phrase	Alteration	Gesture Type	Begin Time	Interruption Time	End Time
1	CR	we're	ah I mean	we're	PM	5.80	6.16	7.48
2	CM	is to	um	is	4	40.80	41.93	43.60
3	CM	there are	um	there's	2	43.34	44.39	46.47
4	CR	which is	um	which is	0	77.53	77.97	80.12
5	CM	in front of that		in front of the two houses	1	84.81	85.55	86.92
6	FS	we could		if one of us goes	2	92.26	92.72	94.32
7	CM	front of the moo	eh ah	front of movie	3	94.94	95.55	97.13
8	CM	then	um	and	1	97.98	98.46	99.13
9	CM	we could be	um	some other people could be	4	102.68	103.22	104.49
10	CM	but what		so what	1	108.09	108.54	108.87
11	CR	and		and	6	135.65	135.88	136.00
12	CR	I guess	um	I guess	LP	170.33	170.90	173.45
13	CM	call s-		call Jim Carrey's friends	PM	184.58	184.96	189.14
14	CR	it was	um	it was	0	194.44	194.76	196.15
15	CR	the		the	0	198.19	198.37	198.66
16	CM	it was s-		it was a movie	2	208.15	208.55	208.71
17	CM	important		unimportant	2	214.17	214.66	215.77
18	CR	I	yeah	I	0	218.19	218.36	218.63
19	CR	ship them	ah	ship them	0	265.70	266.29	266.80
20	CM	I guess we could		I guess we just	PM	296.75	297.54	298.44
21	CR	in your		in your	PM	309.40	309.67	310.94
22	CM	I w-		I'll sure	0	313.01	313.42	313.87
23	CR	but		but	0	316.89	317.27	318.27
24	CM	I		I'm	4	324.46	324.71	325.04
25	FS	I'm		they should	2	324.71	325.04	325.71
26	CM	I s-		I bet	0	380.07	380.23	380.64
27	CR	where		where	0	383.99	384.25	384.72
28	CR	or		or	0	404.21	404.50	405.47
29	CM	maybe we could just cover the the entire	like	using helicopters we could cover the entire	0	407.10	408.50	411.12
30	CR	we	so	we	PM	475.04	475.30	476.44
31	CR	I'm gonna make you eat		I'm gonna make you eat	0	492.69	493.36	494.18
32	CR	I bet		I bet	0	526.53	526.85	527.17
33	CM	wa-		other	0	541.81	541.96	542.26
34	CR	I like the		I like the	0	586.30	587.08	587.60
35	CR	we'll		we'll	0	595.26	595.40	595.51
36	FS	I don't unders-	I mean	what if the wombats	1	596.02	596.75	758.44
37	CM	we could	ah	we	PM	624.21	624.48	625.28
38	CR	we	yeah	we	0	625.21	625.28	625.56
39	CR	first		first	0	661.78	662.39	663.14
40	FS	we just		I go	0	667.68	668.08	668.79
41	CM	with the net		with the nets	PM	671.34	671.88	672.78
42	CM	and then if		and after that	2	674.33	675.02	676.20
43	CR	and		and	0	694.97	695.32	696.37
44	CM	no we ha-		no	0	750.23	750.73	751.40

Table 6. A listing of the speech repairs associated with wd11-spk1. Repair type indicates whether the repair is a content repetition (CR), content modification (CM), or false start (FS). The beginning and ending time of the speech repair, including the reparandum, editing phrase, and alteration, is indicated along with the time of interruption at the end of the reparandum. Note that LP stands for long pause and PM for pragmatic movement.