Fishing or a Z?: Investigating the Effects of Error on Mimetic and Alphabet Device-based Gesture Interaction

Abdallah El Ali
ISLA, University of Amsterdam
Amsterdam, The Netherlands
a.elali@uva.nl

Johan Kildal
Nokia Research Center
Helsinki, Finland
johan.kildal@nokia.com

Vuokko Lantz
Nokia Research Center
Helsinki, Finland
vuokko.lantz@nokia.com

ABSTRACT

While gesture taxonomies provide a classification of device-based gestures in terms of communicative intent, little work has addressed the usability differences in manually performing these gestures. In this primarily qualitative study, we investigate how two sets of iconic gestures that vary in familiarity, mimetic and alphabetic, are affected under varying failed recognition error rates (0-20%, 20-40%, 40-60%). Drawing on experiment logs, video observations, subjects' feedback, and a subjective workload assessment questionnaire, results revealed two main findings: a) mimetic gestures tend to evolve into diverse variations (within the activities they mimic) under high error rates, while alphabet gestures tend to become more rigid and structured and b) mimetic gestures were tolerated under recognition error rates of up to 40%, while alphabet gestures incur significant overall workload with up to only 20% error rates. Thus, while alphabet gestures are more robust to recognition errors in keeping their signature, mimetic gestures are more robust to recognition errors from a usability and user experience standpoint, and thus better suited for inclusion into mainstream device-based gesture interaction with mobile phones.

Categories and Subject Descriptors

H.5.2 [**User Interfaces**]: Input Devices and Strategies

General Terms

Experimentation, Design, Human Factors

Keywords

Device-based gesture interaction, mimetic gestures, alphabet gestures, errors, workload, usability

1. INTRODUCTION & MOTIVATION

Whether we like it or not, errors and failures are an inevitable part of interaction with technology. Device-based gestures (gesturing by moving a device in 3-dimensional space), used in research settings, home environments, or as part of everyday mobile interactions, are gaining potential in becoming a suitable alternative to keyboard or touchscreen-based input, especially when users are encumbered (e.g., manual multitasking [12]). Yet whether a gesture fails due to poor system design or due to the user's actions, errors impede smooth (multimodal) interaction, as well as the adoption of these novel gesture-based interaction methods. This is critical if gesture-based interaction is included in consumer mobile phones.

These device-based gestures (or motion gestures [20])[1] are to be distinguished from surface 'touchscreen' gestures [22]), which typically involve two-dimensional gesture interaction on a surface, such as a tabletop or mobile touchscreen. In contrast, device-based gestures typically make use of accelerometer and gyroscope sensors to allow users to rotate the device in the air for performing gestures in three-dimensional space. While there has been extensive taxonomy-driven work on classifying gestures by referential function into classes [20, 19] and recently identifying usable sets of gestures as defined by users [20, 18, 10], it is still an open question which set of gestures are most robust to errors in gesture-based interaction from a performance-centered standpoint.

In this paper, we look closely at device-based gesture performance using two iconic gesture sets, mimetic (e.g., mimicking a handshaking behavior while holding a mobile device) and alphabet gestures (e.g., drawing the letter 'N' in the air using a mobile phone as a brush), and investigate how their referent-independent performance is affected by varying failed recognition rates.

1.1 Research Questions

Our main research question is: how do mimetic and alphabet gesture sets evolve in the course of interaction when the performed gesture is not recognized under varying error rates? Specifically, we investigate how users react when they are less or more familiar with the ideal shape of a gesture, under varying error conditions. Our hypothesis is that since mimetic gestures are less familiar than alphabets, we expect participants to call on their own real-world sensorimotor experience (on how they perform certain activities in daily life) to perform a gesture. We expect the performance of a mimetic gesture to be influenced by this experience, thus morphing the iconic gesture into that previously learned gesture. For example, while a 'fishing' gesture might be designed in a particular way to perform some system function (e.g., hold device flat on palm, tilt towards you, and place device back on palm), given unfamiliarity with its designed ideal form, we suspect that this same gesture is more likely to morph into a more natural, learned fishing gesture upon failed recognition attempts. Given the many ways to fish (where variations could be due to cultural or individual differ-

[1]We use the term 'device-based gestures' and not 'motion gestures' to emphasize that these gestures require holding a device, and not solely about motion as in vision-based gesture recognition such as interacting with a Microsoft Kinect©.

ICMI'12, October 22–26, 2012, Santa Monica, California, USA.

ences), we expect the evolved gesture to exhibit more variation in the face of increasing error, especially since subjects do not know the ideal shape of the mimetic gesture. Complementarily, this variation arises due to the higher degrees of freedom permitted in performing that gesture.

By contrast, we expect alphabet gestures to exhibit much less variation, instead becoming more rigid and structured after repeated failed recognition attempts – this is because alphabet gestures not only have lower degrees of freedom (namely, 2df), but the set of ideal visual shapes is more familiar to users. Similar to work in speech and handwriting recognition, we expect a hyperarticulation [16, 14] of gestures as subjects begin to gesture more carefully to arrive at the ideal shape required by the recognizer. This we hypothesize will negatively impact the user experience (UX)[2] of performing these gestures.

Investigating the usability differences in a primarily qualitative manner between mimetic and alphabet gestures here yields two main research contributions: first, it aids gesture designers in selecting which gesture set (mimetic or alphabet) is more robust to errors, and hence better suited for inclusion into accelerometer and/or gyroscope equipped mobile devices. This is achieved by providing a deeper understanding of whether some gestures are intrinsically more tolerant to recognition errors than others (e.g., by showing that mimetic gestures, in lacking an ideal shape, can foster more variation in gesture-based interaction). Second, it equips gesture designers with the knowledge of which gesture sets overall induce a lower subjective workload to perform, especially under conditions of high recognition failure.

Additionally, we provide initial results on how errors can be an impeding factor in performing gestures in public, as well as use-cases for the tested gestures subjects reported on. The rest of the paper is structured as follows: first we provide a review of related work, then we present our study design and methods, give our results and discuss them, and finally conclude and hint at future work.

2. RELATED WORK

2.1 Gesture-based Interaction

Recent work has looked into the user preferences of certain gestures given a task (e.g., call answering), where the goal was to arrive at a taxonomy of gesture-task pairs that can aid device-based gesture design [20]. Relatedly, a recent study [5] has investigated the naturalness and intuitiveness of gestures, where the goal was to understand how users' mental models are aligned to certain gestures. Another line of research has focused on the social acceptability of produced gestures under different settings (e.g., at home, at the pub, etc.) [18]. The goal here was to equip gesture designers with knowledge of which gestures are socially appropriate under which settings and situations. While much research has focused on the naturalness, intuitiveness, and the social consequences of performing certain (surface and device-based) gestures, little research has addressed how issues of failed recognition can transform a produced device-based gesture in the course of interaction.

2.2 Dealing with Recognition Errors Across Modalities

Human factors research in multimodal interaction concerned with recognition errors [11] is a well researched topic in multimodal interfaces [2, 13, 15], where investigations were typically concerned with error handling strategies devised by users in the face of recognition errors (e.g., modality switching to a 'familiar', more efficient modality). In speech-based interfaces, a common finding is that the most intuitive and instinctive way for correcting errors in speech is to repeat the spoken utterance [21] and hyperarticulate it [14]. For multimodal systems, a common error-correction strategy is to repeat a modal action at least once prior to switching to another modality [15]. In [15], they observed a repetition count (called 'spiral depth') of depth 6, before users would switch to another modality. In a follow-up study by [6], they tested 3 commercial Automatic Speech Recognition (ASR) systems where they found that a little over 50% of the time, subjects would continue to repeat the utterance to a spiral depth of level 3.

However, while recognition errors have been well studied in domains such as speech-based interfaces [2, 14], handwriting recognition [16, 17], and multimodal interfaces [13], less attention has been given to usability issues surrounding device-based gesture interaction. An exception is the study by [8], where they investigated user tolerance for errors in touch-less computer vision-enabled gesture interaction under both desktop (keyboard readily available in front of subjects) and ubiquitous computing settings (keyboard not readily available). In this Wizard-of-Oz study, they found that the interaction context played a significant role in how tolerant users were to errors. Specifically, they found that in the ubiquitous computing scenario, users are much more tolerant to errors than in the desktop condition (where recognition error rates can potentially reach 40%) before users will abandon gesture based interaction in favor of traditional, keyboard-based input.

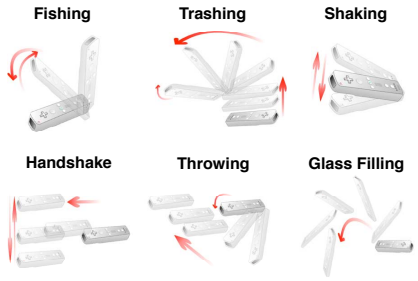
While the results of the study by [8] are relevant to the current work, our work differs in three main ways: first, their goal was to investigate whether and to what extent subjects would switch modalities when confronted with recognition errors, and not study how gestures evolve in response to error. Second, they were concerned with a single gesture, and not different gesture sets and their respective performance by users under different recognition error rates. Finally, their concern was with computer vision-enabled interaction, and not device-based gesture interaction.

2.3 Gesture Taxonomies

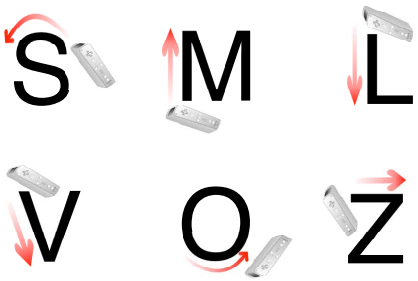
In [19], they distinguish between symbolic gestures and pantomimic gestures. Symbolic gestures are gestures that have come to take up a single, culturally-specific meaning. Examples include American Sign Language (ASL) gestures and natural language alphabets (e.g., letter 'C'). Pantomimic gestures on the other hand, are used for showing (through mimicry) the use of movement of some invisible tool or object in the speaker's hand. For example, when a speaker says "I turned the steering wheel hard to the left", while mimicking the action of turning a wheel with both hands, they are performing a pantomimic gesture. Here, we are not concerned with the external gesture referent, only its iconic movement in space.

In [20], in order to study user-defined motion gestures, they devised an initial taxonomy by which to classify gestures. Here, they distinguished between metaphor gestures (metaphor of acting on a physical object other than the phone), physical gestures (direct manipulation), symbolic gestures (visually depicting a symbol by e.g., drawing with the phone), and abstract (gesture-activity mapping is completely arbitrary). In the foregoing taxonomies, the classifications are based on the communicative intent behind performing the gesture (i.e., its representational aspect), and not on the usability and user experience (UX) of manually performing the gesture.

[2]UX here is based on ISO 9241-210 [1] definition: "A person's perceptions and responses that result from the use or anticipated use of a product, system or service."



(a) Mimetic gestures.



(b) Alphabet gestures.

Figure 1: The designed gesture sets for (a) mimetic gestures and (b) alphabet gestures.

3. METHODS

3.1 Study Design

In this study, we do not to test the communicative intentions of subjects while performing gestures (i.e., no referential aspect), but only the manual 'performance' of the iconic gesture itself. Under this performance-centric view, the mimetic gestures are iconic movements of real-world actions and the alphabet gestures are iconic movements of writing letters.

Mimetic gestures were chosen as they provide a natural means of interaction, making them good candidates for inclusion into mobile technology. Similarly, alphabets can be produced by any literate person, and hence suitable for comparison. Other symbols (e.g., salute or high-five gesture) were not tested to avoid undue cultural bias. Furthermore, like mimetic gestures, alphabets can also be easily learned and recalled, and have practical potential for use in mobile technology (e.g., mid-air writing for user authentication [9]). Given these design considerations, we designed 12 gestures (6 mimetic, 6 alphabetic). Mimetic gestures are: Fishing, Trashing, Shaking, Handshake, Throwing, Glass Filling. These specific gestures were chosen because they represent familiar yet different activities, where all have more than 2 degrees of freedom. Alphabet gestures we chose are English letters (given our subject pools' language), varied by difficulty (e.g., 2 strokes or 3 to draw the letter). Only 6 different gestures for each set was chosen to avoid participant fatigue, given the high number of trials (200 total) each participant entered. Nevertheless, we believe the chosen gesture samples provide sufficient variation to characterize each gesture set with respect to varying failed recognition rates. Both gesture sets, and their movement in space, are shown in Fig. 1.

To investigate how different types of performed gestures respond to varied recognition errors, we used an automated Wizard-of-Oz method [4] where we simulated each of three failed recognition error rate conditions: low (0-20%) error rate, medium error rate (20-40%), high error rate (40-60%). Subjects were told that real gesture recognition engines were being tested, where each of the algorithms differs in terms of how sensitive it is to the user's gesture interaction style. When subjects performed a gesture, the automated wizard would draw for each gesture block an error rate randomly from the assigned error rate range specific to the condition. When a gesture is performed, the subject receives coarse feedback (recognized / not recognized) on whether or not the performed gesture was successfully recognized.

For this study, testing real recognizers is irrelevant as we are only interested in the usability and user experience (UX) of gesture performance under the chosen error rates. This is both in line with previous work [8], and conveniently allows testing our research questions without the unpredictability of real recognizers. Importantly, here we study gesture performance, not how different feedback improves gesture learnability. Additionally, we tested task-independent gestures for two reasons: first, it allows us to understand the differences between the two gesture sets (mimetic and alphabet) independent of task type. This would eliminate potential subject bias in both workload and expected gesture evolution under error conditions due to the mapping of a given gesture to a task. Second, following [10], it allows subjects to freely speculate about the applied real-world use of these two gesture types.

The conducted experiment was a mixed between- and within-subject factorial (2 x 3) design. A between-subjects design between the two gesture sets was chosen for two reasons: first, to disallow any transfer effects between the two gesture sets thereby avoiding any contamination between the gesture mental models formed in subjects. Second, testing all gestures in one session would excessively lengthen the duration of the experiment, and pose a risk of participant fatigue. There are two independent variables (IVs): gesture type (2 levels: mimetic vs. alphabet) and recognition errors (3 levels: low (0-20%) vs. medium error (20-40%) vs. high error (40-60%), where gesture-type was a between-subjects factor and error rate a within-subjects factor. Each between-subject condition tested 6 gestures (12 total), randomized across subjects. Each gesture occurred in all within-subject conditions (counterbalanced across subjects), in addition to two practice blocks, which resulted in 20 gesture blocks per experimental session. Each block consisted of 10 trials. In a block, subjects were asked to perform a given gesture using a Wii Remote® 10 different times (once per trial), where the error rates are randomly distributed within the corresponding recognition error level. In the practice blocks however, the error rate was always low. In total, each subject entered 200 trials (20 practice, 180 test).

The experiment was coded using NBS Presentation®[3], an experimental control and stimulus delivery software. Interaction and syncing with the Wii Remote was done using GlovePie[4], a programmable input emulator. Four data sources were collected: modified NASA-TLX workload data [3, 7], experiment logs, gesture video recordings, and post-experiment interviews. The modified NASA-TLX questionnaire assessed participants' subjective workload quantitatively ([0,20] response range) through the index's constituent categories: Mental Workload, Physical Workload, Time Pressure, Effort Expended, Performance Level Achieved and Frustration Experienced [7], plus the additional categories of Annoyance and Overall Preference [3]. Given no time pressure imposed on subjects in the study, we did not use this category. Additionally,

[3]http://www.neurobs.com/; last retrieved: 15-08-2012

[4]http://sites.google.com/site/carlkenner/glovepie; last retrieved: 15-08-2012

as the Annoyance category is specific to audio interfaces, we only made the additional use of the Overall Preference category.

3.2 Subjects

24 subjects (16 male, 8 female) aged between 22-41 (M_{age}= 29.6; SD_{age}= 4.5) were recruited. Our subject sample spanned 8 different nationalities, where all but one were right-handed (23/24). Many subjects (17/24) had a technical background, and most (19/24) were familiar with gaming consoles that use some form of gesture recognition technology (e.g., Nintendo Wii© or Microsoft Kinect©).

3.3 Setup & Procedure

The experiment was carried out at the usability lab at XYZ center. Each experimental session took approximately 1 hour to complete. Subjects were seated in front of a monitor, where a tripod-mounted camera was aimed at their gesture interaction space. They were allowed to define their own interaction space to ensure their comfort during the session, so long as it was still within the camera's view. Prior to the experiment, each subject filled a background information form, signed an informed consent form, and read through detailed instructions for performing the task. After reading the instructions, a tutorial was given on how to perform each gesture. The tutorial involved the experimenter performing each gesture (using the Wii Remote) right next to the subject.

The first two blocks were practice blocks, set always at a low error rate. Before each block, a video of how the gesture to be performed in the next trials was shown on the screen. The performance of the gestures in the videos was identical to how they were performed by the experimenter in the tutorial. The videos were shown to eliminate any failed memory recall effects, where subjects' (multimodal) interaction requires a visual input (videos watched) and a translation to somatosensory output (performed gesture). In a trial, the subject would be instructed on screen in text to perform the gesture for that block (e.g., "Perform Fishing gesture."). A subject would have to press and hold the A button on the Wii Remote to start recording the gesture, and release it after completing the gesture. After performing the instructed gesture, if the subject falls into a successful gesture recognition trial, a green checkmark image is flashed on the screen, while in a failed recognition trial, a red image with the word "FAIL" is flashed on the screen. After each block, subjects were asked to fill in the modified NASA-TLX questionnaire, where they were provided with an optional 2 min. break after completing it. Subjects were allowed to change their responses on the questionnaire at the end of each block, so that their responses per block can be truly relative to one another. After completing the experiment, subjects were interviewed for around 10 min. about their experience with the experiment. Afterward, they were thanked for participating, and offered a movie theatre ticket.

4. RESULTS

4.1 Subjective Workload

The modified NASA-TLX responses were analyzed within groups, per type of gesture. For each modified NASA-TLX category, one-way ANOVA repeated measures tests were conducted comparing results from all error rates. A mixed between- and within-subjects ANOVA revealed no significant differences between the two gesture sets, and therefore not reported. Descriptive statistics (means, standard deviations, 95% confidence intervals, p-values, partial eta squared) of within-subject results for the mimetic and alphabet gestures are shown in Table 1 and Table 2, respectively. Means and confidence intervals for each category under mimetic and alphabet gesture conditions are shown in Fig. 2 and Fig. 3, respectively. Post-hoc pairwise comparisons (with Bonferroni correction[5]) between error conditions (Low-Medium, Low-High, Medium-High) were conducted in every case. Where significant, they are represented in the graphs as bars between low and medium error rate conditions, and as brackets between low and high error rates, with the corresponding significance levels (**, $p < 0.01$; *, $p < 0.05$).

Mimetic Gestures

Category	Error	M	SD	95% CI	P-value	η_p^2
Mental	Low	5.3	3	[3.6, 7]	p= .098	
	Med	5.8	3.3	[4, 7.7]	$F(2,22) = 2.6$	.2
	High	7.2	3.5	[5.3, 9.2]		
Physical	Low	8	4.9	[5.3, 10.8]	p = .365	
	Med	7.7	4.5	[5.1, 10.2]	$F(2,22) = 1$	.1
	High	8.4	5	[5.6, 11.3]		
Effort	Low	7.5	4.7	[4.8, 10.2]	p=.119	
	Med	7.7	4.4	[5.2, 10.3]	$F(2,22) = 2.3$	.2
	High	8.9	4.5	[6.4, 11.5]		
Perform.	Low	15.6	2.6	[14.1, 17.1]	**p<.05**	
	Med	13.7	1.9	[12.6, 14.7]	$F(1.4,15.1) = 11.2$	.5
	High	10.3	4.4	[7.8, 12.8]	(corr. G-G ϵ= .68)	
Frust.	Low	5.5	3.1	[3.7, 7.3]	**p<.001**	
	Med	7.7	3.6	[5.7, 9.7]	$F(2,22) = 23.2$	.7
	High	10.7	4.3	[8.3, 13.2]		
Pref.	Low	13.3	2.7	[11.8, 14.9]	**p<.05**	
	Med	11	2.6	[9.5, 12.5]	$F(2,22) = 5.1$	.3
	High	9.7	4	[7.5, 12]		
Workload	Low	5.1	2.1	[3.9, 6.3]	**p<.001**	
	Med	5.9	2.2	[4.6, 7.2]	$F(2,28) = 13.2$	.5
	High	7.5	2.8	[6, 9.1]		

Table 1: Descriptive statistics for mimetic gestures (N=12) under different error rates (Low, Medium, High).

Alphabet Gestures

Category	Error	M	SD	95% CI	P-value	η_p^2
Mental	Low	5.8	3.1	[4, 7.6]	p=.872	
	Med	5.5	2.7	[3.9, 7.1]	$F(2,22) = .1$	.01
	High	6	2.2	[4.8, 7.2]		
Physical	Low	10.2	3.4	[8.3, 12.1]	p = .403	
	Med	9.2	3.4	[7.3, 11.2]	$F(1.3,14.1) = .9$	.08
	High	10.8	3.2	[9, 12.7]	(corr. G-G ϵ= .69)	
Effort	Low	7.4	2.4	[6, 8.8]	**p<.05**	
	Med	8.8	2.9	[7.1, 10.4]	$F(2,22) = 6.8$	.4
	High	9.9	3.5	[7.9, 11.9]		
Perform.	Low	15.5	2.2	[14.2, 16.8]	**p<.001**	
	Med	12.3	2.4	[11, 13.7]	$F(1.2,13.5) = 23.5$	.7
	High	8.7	4.7	[6.1, 11.4]	(corr. G-G ϵ= .61)	
Frust.	Low	4.2	2.5	[2.8, 5.7]	**p<.001**	
	Med	6.3	3.2	[4.5, 8.2]	$F(1.1,12.6) = 11$	.5
	High	9.6	5.6	[6.4, 12.8]	(corr. G-G ϵ= .57)	
Pref.	Low	13.3	2.5	[11.9, 14.8]	**p<.05**	
	Med	11.2	2.8	[9.7, 12.8]	$F(1.3,14.3) = 23.6$	.7
	High	8.3	3	[6.6, 10]	(corr. G-G ϵ= .65)	
Workload	Low	5.4	1.1	[4.7, 6]	**p<.001**	
	Med	6.2	1	[5.7, 6.8]	$F(2,22) = 16.5$	.6
	High	7.9	2.1	[6.7, 9.1]		

Table 2: Descriptive statistics for alphabet gestures (N=12) under different error rates (Low, Medium, High).

Since there were no significant differences between the two gesture sets, it appears to be that experimentally our wizard-of-Oz recognizer for both gesture conditions had a similar effect on participants, where statistically the two independent groups did not treat the gesture sets differently. However, there were differences between gesture sets with respect to the differences in error conditions. Based on the within-subjects ANOVA results and post-

[5]Backward-corrected SPSS© Bonferroni adjusted p-values are reported.

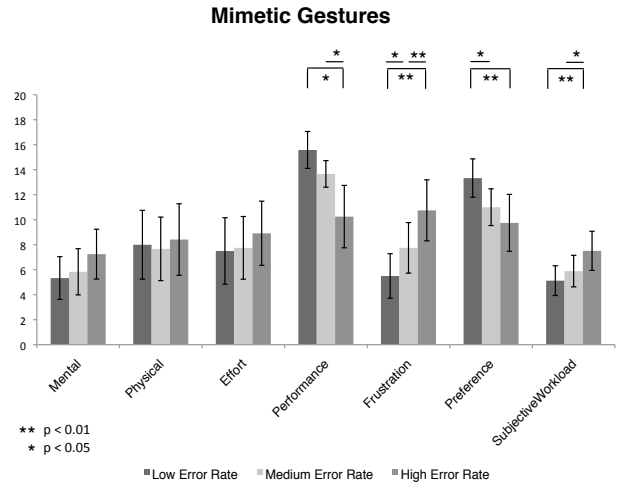


Figure 2: Modified NASA-TLX workload measurements for mimetic gestures. Capped error bars represent a 95% confidence interval.

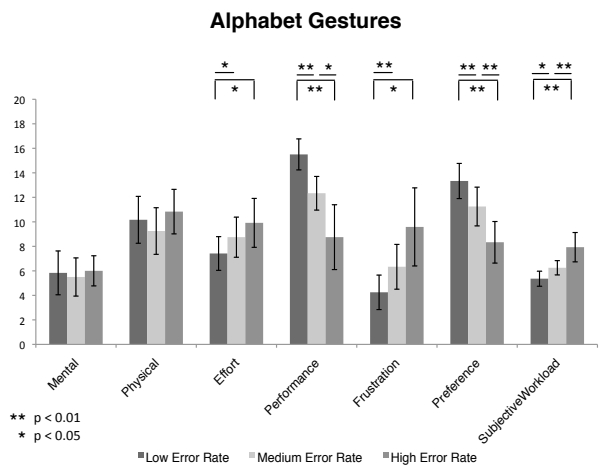


Figure 3: Modified NASA-TLX workload measurements for alphabet gestures. Capped error bars represent a 95% confidence interval.

hoc pairwise comparisons on the modified NASA-TLX scores for mimetic gestures (Table 1; Fig 2) and alphabet gestures (Table 2; Fig 3), we summarize our findings.

Results showed that while subjects in the alphabet gesture condition had to place significantly more effort (Effort Expended) between the low error rate and the medium and low and high error rate conditions, this was not so for the mimetic gestures. Additionally, for the mimetic gestures, Performance Level Achieved between the low and medium error rates was not significant, while it was significant across all error rates for the alphabet gestures. In the mimetic condition, Frustration Experienced was significant across error conditions, however in alphabet gestures, Frustration was significant only between Low-Medium and Low-High. This shows that for alphabet gestures, frustration is more or less consistently experienced beyond 20% error rates. Interestingly, Overall Preference in the mimetic gesture condition fell significantly only between the low and high error rates and between low and medium, while it significantly dropped for each error rate in the alphabet gesture condition. This hints at a feeling of helplessness in the face of errors for alphabet gestures, possibly because fewer parameters can vary for this gesture set when subjects repeatedly try to recover from these errors and still fail.

Finally, Subjective Workload for subjects in the mimetic gesture condition significantly differed only between low and high error rate conditions, and between the medium and high error rate conditions, while there were significant workload increases across all error rates for subjects in the alphabet gesture condition. Together, these findings suggest that mimetic gestures are better tolerated under error rates of up to 40% (cf., [8]), compared with error rates of up to only 20% for alphabet gestures. From a usability perspective, our workload results indicate that mimetic gestures are more robust to recognition failure than alphabet gestures, likely due to the higher design space available for users to experiment with given their unfamiliarity with the ideal shape of the designed gesture.

4.2 Video Analysis

Observation of subjects' behavior from the videos provided early findings on how mimetic and alphabet gestures evolve under varying error conditions.

Mimetic vs. Alphabet Gesture Evolution. While evolution in gesture performance for both sets was observed in as few as 2 successive error trials (i.e., spiral depth of 2), less variation in gesture performance was observed in the alphabet gesture condition. In both conditions however, we observed that the continued successful recognition of a gesture served as a continuos reinforcer for repeating a gesture in the same manner, irrespective of gesturing style. If however a subject chose to experiment with a variation during those success trials, then the variation was repeated on grounds that it will work on a subsequent trial. We call this repeated gesture variation the 'canonical variation'. As shown below, these observations are corroborated by subjects' feedback during the interviews.

We observed the following canonical variations of mimetic gestures under high error: S4's Handshake canonical variant was extending his arm straight and swinging the Wii Remote up and down as if it is a real hand he is reaching out to. S8 exhibited a variation in the Trashing gesture, where the speed of the gesture increased drastically and the end position was raised higher than shown in the video. S10 arrived at the canonical variation of the Glass Filling gesture, which required a slow, calculated twisting of the wrist while the Wii Remote was in the pouring position. S12 exhibited variations on both Fishing and Trashing gestures: Fishing gesture resembled real-world fishing, which involved lightly pulling the Wii Remote diagonally upwards and then with slight force throwing the Wii Remote back as if into an imaginary water pool. The Trashing variation was a trashing gesture that was both small in breadth and with almost no movement of the arm and shoulders.

For alphabet gestures, apart from brief intermittent experimentation with breadth, speed, position, completeness and direction of gestures, the main systematic variations observed were attempts to draw the letter symbol more precisely under high error conditions (cf., [17]). However, there was one subject who appeared to have arrived at a canonical variation: the letter Z was drawn very quickly with the last stroke (bottom line of a 'Z') more slanted.

Persistence vs. Evolution. When do recognition failures cause subjects to persist in repeating the same gesture, and when to push them to explore new gestures? From our observations, certain patterns emerged: first, it seems that if a subject performs a gesture quickly, and it fails, he will experiment with a slower version. If he first performs it slowly, then he will experiment with a faster version. Second, after repeated success trials, the speed of the ges-

ture is the first parameter to vary while the other parameters mostly remain constant, irrespective of gesture class. Usually, though not always, only after failure does exploration take place, where variations come into play. Third, if in a block subjects experience a series of successes (4-5), they will be more likely to repeat the same gesture even in the face of repeated errors later in the block, irrespective of gesture class. This successive positive reinforcement suggests that subjects have figured out what the recognizer wants in that block. Additionally, if a gesture succeeds too many times in succession ($\geq$4), people seem to apply the principle of least effort and perform incomplete gestures (as witnessed by a downsized version of the Throwing gesture by S6 in the low error condition). This was observed mainly in the alphabet gesture condition, where two subjects later mentioned in the interviews that they did not need to complete the gesture for recognition to take place.

4.3 Users' Subjective Feedback

Perceived Canonical Variations. Many subjects (18/24) across both gesture sets reported that in the face of repeated errors, they would start experimenting with different variations of the gesture, and when feedback was positive, they replicated that variation. This suggests that positive reinforcement after repeated error trials was the driving force behind the step-wise evolution of a given gesture, or put differently, survival of the fittest gesture variation. Supporting our observations, subjects reported much less variation in how they performed gestures under high error rate conditions in the alphabet gesture condition than in the mimetic condition. For mimetic gestures, subjects reported many novel strategies for how and when a given gesture was recognized (*S9: "The shaking, that was the hardest one because you couldn't just shake freely [gestures in hand], it had to be more precise shaking [swing to the left, swing to the right]... so not just any sort of shaking [shakes hand in many dimensions]"; S11: "What I noticed was that the fishing for example, it was just a rotation [small upward wrist rotation], and when you did it as if you were really fishing, then it didn't work*.").

In line with our hypothesis, the variations under the alphabet gesture condition were perceived to involve a more precise and well structured gesture for recognition to be successful (*S17:"I think within certain blocks I got the pattern of what was working. It was more apparent in the third one [high error rate condition], you have to do it better.*"). There were exceptions to this, mainly the drawing of the letter O, which can easily vary (in addition to breadth and speed) along the direction and the start/end position parameters (*S16: "I didn't really change the way I did them [the gestures], except for the O, to see different ways it can be written*."). However, while variations that involved more rigid gesturing of letters was both observed and reported by subjects, some subjects explicitly expressed arriving at the canonical variations of some letters (*S22: "In Z or S, I made them more rounded, and then it worked better this way.*").

Individual and Cultural Differences. For the mimetic gestures, subjects found the Throwing and Glass Filling gestures the most problematic (4 reports each). Throwing in particular appeared not to be as intuitive as the other gestures, possibly because of the many ways people can throw (*S7: "I would have done the throwing gesture differently [shows experimenter how different people throw]*") The mapping to real-world behaviors was evident when explaining why the Glass Filling gesture is difficult (*S10: "For the glass filling, there are many ways to do it. Sometimes very fast, sometimes slow like beer.*") Interestingly, the Shaking gesture, which is already an input technique in some smart phones (e.g., Apple's iPhone 4®[6]), was not taken favorably by at least one subject: *S4: "Shaking was hard... how long should you shake? I tried here, and it was enough to do two movements... and you have to use your wrist, which can be a problem*.").

Two subjects noted that the Fishing gesture was quite easy to perform, however the naturalness of the gesture backfired (*S10: "In the fishing gesture, I was just guessing how to do it. Because I have never done it practically. I cannot really see myself performing well, just simulating it*.") Likewise for the Trashing gesture: (*S10: "I was trying to emulate how I would normally do it. For example trashing, sometimes you're not in the mood, and you do it like this [trashes imaginary object downward softly], very quietly.*") Together, these reports support our hypothesis that mimetic gesture evolve into their real-world counterparts, especially when under high error conditions. Due to the importance of cultural differences in performing certain mimetic gestures, it would be interesting to see whether these cultural forms are the first gestures subjects recourse back to under error conditions.

For the alphabet gestures, the O and M alphabet gestures were perceived to be the most difficult (6 and 5 votes, respectively). The O gesture was perceived to be difficult on grounds that there are many ways to write/draw an O (*S17: "The O was a bit funny, because naturally I start from the top, not the bottom*."). This was likely a fault of how the videos in the experiment showed the gesturing of the O, which began from bottom to top. However, the other reason for finding the O difficult was due to the completeness and position parameters (*S13: "For the O, I noticed that if I start here [points in space] and I end here, and there is a gap, then it wouldn't be recognized*."). There were no clear reasons given for why the M gesture was difficult, other than hinting at the number of strokes required (*S14: "The M is more articulated, so you have to spend more time on it*."). V was perceived to be the easiest letter to gesture, due to the ease by which it can be drawn. This was so despite that V, like the O, varied along the direction parameter.

Perceived Performance. In general, subjects reported they were pleased with their overall performance in both mimetic (7/12) and alphabet (7/12) gesture conditions. Surprisingly, while all subjects noticed the difference in difficulty across blocks, some subjects in the mimetic gesture condition (6/12) and some in the alphabet gesture condition (7/12) seemed to treat their performance on the low error rate and the medium error rate conditions as more or less the same (*S18: "Between the first and second blocks [low and medium error rate conditions], it was the same... 10-15%*"), while attributing very poor performance to the high error rate condition. The reason for this was likely due to the extremely high error rate range (40-60%). This relates to the question of how poor can performance of gesture recognition technology get before the technology is abandoned in favor of well-established methods such as keyboard-based input (cf., the 40% error threshold set by [8]).

Additionally, subjects showed an incredible ability to justify their performance. If subjects fell into the high error rate condition first, they attributed their poor performance to the fact that they are still learning how to perform the gestures (*S22: "In the beginning, it was less because I was still learning*."). By contrast, if the high error condition comes later in the block, they attribute their poor performance later due to not putting the kind of effort and attention they did on the first block, which supposedly explains their better performance earlier (*S7: "For the third block [high error rate condition]...I had done many already, I was like I'll just do it and see what happens*."). This is in line with our video observations, where

[6]http://www.apple.com/iphone/; last retrieved: 15-08-2012

gesture performance was different in high error conditions where gestures tended to evolve more for mimetic gestures (poor observed performance) and get more rigid for alphabet gestures (acceptable observed performance if canonical variation is right).

Use of Gestures on Mobile Phones. For mimetic gestures, when asked whether they saw any real-world use of the tested gestures on mobile phones, most subjects (10/12) reported at least one use-case. For the Throwing gesture, a mapping to sending a message was identified as a reasonable interaction method. Another example of Throwing was mobile payment, where one could throw money to a cashier. Similarly for the Handshake gesture, where the handshake could be a form of digital content transaction or a form of business e-card exchange. Trashing was implicated in hanging up a call, deleting content, or what subjects did not favor, to turn off an alarm clock by turning over the device. Subjects reported that shaking was already available in some mobile phones used for switching to the next song. However, some subjects expressed that some of the gestures simply had no use (*S4: "Trashing at a conference is quite natural to turn off a call, but handshake, I can't think of a use*.").

When asked about using alphabet gestures on mobile phones (e.g., gesturing 'S' for sending a message), only half of subjects said they would use such gestures (6/12) (*S24: "If there was a phone with this, I would really like that!*"), and some thought it depended on the situation (3/12) (*S19: "If you're on your bike, you could do that*."). From those who would use such gestures, they explicitly mentioned that they have to necessarily be free of error (*S20: "Yeah [I would use such a gesture], if it's really foolproof*.";). Reported use cases included gesturing P to pause a game, C for calling or checking a calendar, or most reported, for traversing an interface's menu structure (*S18: "If you need to access a feature that is hidden away in the menu, like starting to write an SMS, then gesture S*"). Some subjects mentioned that the gestures need to be explicitly different (*S22: "Those gestures need to be so different so you cannot do wrong gestures...V is part of M*.") and others required a distinct gesture-to-task mapping for each application on your device ("*S23: "The applications should not clash for the same gesture*."). Finally, one subject remarked that while these alphabet gestures might work for the roman alphabet, it would be radically different for chinese characters.

Social Acceptability. For both the mimetic and alphabet gestures, some subjects (8/24) expressed concern over performing these gestures in public. This is in line with previous work, which explicitly addressed the question of what kinds of gestures people would perform while in public [18]. It was surprising to find that there were very few concerns about performing mimetic gestures in public (2/12) (*S9: "I'm not sure [about Fishing], because I'm not fishing myself, why am I doing this?*"), as opposed to performing alphabet gestures (6/12) (*S13: "If I am not in a public place, then yes. Otherwise, you would think I'm a Harry Potter in disguise!*").

While recognition errors play a clear role in preventing gesturing in public settings (*S16: "When it doesn't take your C, you keep doing it, and it looks ridiculous*."), the breadth of a gesture was also perceived as an important factor (*S18: "Maybe if I do it small, if I don't look very freaky, then it's okay*."). Still, others thought it fun to try novel things in public (*S14: "I never thought about that [alphabet gestures], but why not? I would not find it embarrassing*.").

5. DISCUSSION

5.1 Limitations

There are three potential limitations to the present study: first, since our study was conducted in a laboratory, it had less ecological validity. While subjects explicitly stated that they would not be comfortable performing some gestures (especially alphabet gestures) in public, from the videos it was evident that all subjects performed all the instructed gestures freely and without hesitation. It is interesting to consider here whether the desire to perform all gestures successfully in each block could have overruled whatever embarrassment might come about from merely performing the gesture to execute some device function. At least one subject explicitly mentioned the importance of performance (*S9: "I eventually got the hang of it, and yeah, I really, really wanted to improve my performance!*"). Additionally, gesture performance under failed recognition rates may not reflect performance when a user is mobile (e.g., walking or in public transport).

Second, while testing how task-independent gestures are affected under varying error rates was an explicit design choice, it could be that a gesture to task mapping is necessary for unraveling the usability of a given gesture. While we agree that the task being performed is an important variable in gesture-based interaction, we nevertheless argue that there are differences between individual gestures and importantly between gesture sets that can be unraveled only by leaving out the task. This is to eliminate any potential bias that the task might evoke. For example, calling someone might be considered more urgent in some situations than skipping to the next song, and that might influence the performance and workload required from a gesture. Finally, our design choice in simulating a gesture recognition engine as realistically as possible meant that the errors had to be randomly distributed in each block. In future work, experimentation with different error distributions would better help understand different types of gesture evolution. Additionally, with a real recognition engine, the precise evolution of the gesturing behavior may differ than what was observed in this study.

5.2 Implications for Gesture Recognition

Our observations and subject reports showed that mimetic gestures and alphabet gestures do indeed differ under increasing recognition error conditions. While our observations and subject reports showed mimetic gestures tend to vary more into their real-world counterparts when they are repeatedly not recognized, alphabet gestures tend to become more rigid and well structured. This is in line with work on other modalities like speech and handwriting recognition [16, 14]. This suggests that for gesture-based interaction to be accepted in the consumer market, accurate recognition from the first attempt appears to be quite important for mimetic gestures. If a gesture is not recognized from the first instance, there is a risk that the subsequent gesture differs radically from the first, which would be beyond the scope of the recognition algorithm. This is in contrast to alphabet gestures, which in having lower degrees of freedom vary in fewer parameters (mainly speed, breadth, and start/end position) under error conditions. This suggests that recognition engines (in uni- or multimodal systems) can more easily deal with post-failure recognition when this set of gestures is used.

Additionally, subjects had no real means to understand the cause of the errors, to avoid errors, or to improve recognition rates. However, they came up with interesting explanations (e.g., canonical variations) why there were more errors in different blocks and what might have caused them (e.g., fatigue). Nevertheless, we observed that they were also active in adapting their gesturing behavior in order to improve recognition errors and to understand the workings of the recognition engine. This seems to suggest that transparency in the gesture recognizer may better support users in their error handling strategies during situations of failed recognition.

5.3 Implications for Gesture-based Interaction

It was evident from our results (modified NASA-TLX workload data, video observations as well as subjects' feedback) that not only do mimetic gestures vary differently than alphabet gestures under error conditions, but also there were differences between individual gestures under each class. We found that while mimetic gestures yield significant increases in overall subjective workload between low and high error rates, and between medium and high error rates, overall workload for alphabet gestures significantly increased across all error rate conditions. This indicates that mimetic gestures are better tolerated under error rates of up to 40%, while alphabet gestures incur significant overall workload with up to only 20% error rates. This is in line with previous work on computer vision-based gesture interaction [8], where our workload results suggested user error tolerance of up to 40% for the mimetic gestures only.

The two gesture sets also differed in potential use and social acceptability. For mimetic gestures, interesting use cases (e.g., handshake for digital content exchange, throwing for mobile payment) were given, while limited use cases (e.g., interface menu structure traversal) were offered for alphabet gestures. Moreover, alphabet gestures were seen as more embarrassing to perform in public, especially if they are not recognized. From a usability perspective, these findings suggest that mimetic gestures are more promising than alphabet gestures for use during device-based gesture interaction, even under conditions of medium recognition error.

6. FUTURE WORK & CONCLUSIONS

A follow-up quantitative study is required to unravel exactly how many errors in succession are required before a given gesture evolves into its canonical variant. While we have presented a qualitative analysis of gesture evolution, quantitative models of the precise evolution behavior would help identify the exact parameter changes across error conditions. Future work will also address how errors influence gesture performance of other gesture sets (e.g., metaphoric and manipulative gestures), and for different device form factors.

In this paper, we described the results of an automated Wizard-of-Oz study to qualitatively investigate how mimetic and alphabet gestures are affected under varying recognition error rates. In line with our hypothesis, it was shown that mimetic gestures, which have a less familiar ideal shape, tend to evolve into diverse real-world variations under high error conditions, while alphabet gestures tend to become more rigid and structured. Furthermore, we showed that mimetic gestures seem to be tolerated under error rates of up to 40% (cf., [8]), while alphabet gestures incur significant overall workload with up to only 20% error rates. From this, we drew usability implications showing the importance of immediate accurate recognition of mimetic gestures (as a way of taming the tendency of these gestures to evolve) and suggested they are better suited than alphabet gestures for inclusion into mainstream device-based gesture interaction with mobile phones.

7. ACKNOWLEDGEMENTS

Authors thank Eurostar (E5262) SmartInside Project for its support.

8. REFERENCES

[1] *ISO 9241-210:2009. Ergonomics of human system interaction - Human-centred design for interactive systems.*

[2] M.-L. Bourguet. Towards a taxonomy of error-handling strategies in recognition-based multi-modal human-computer interfaces. *Signal Process.*, 86:3625–3643, 2006.

[3] S. Brewster. *Providing a structured method for integrating non-speech audio into human-computer interfaces*. PhD thesis, University of York, 1994.

[4] G. D. Fabbrizio, G. Tur, and D. Hakkani-Tur. Automated wizard-of-oz for spoken dialogue systems. In *Proc. INTERSPEECH 2005*, pages 1857–1860, 2005.

[5] S. A. Grandhi, G. Joue, and I. Mittelberg. Understanding naturalness and intuitiveness in gesture production: insights for touchless gestural interfaces. In *Proc. CHI '11*, pages 821–824, NY, USA, 2011. ACM.

[6] C. Halverson, D. Horn, C. Karat, and J. Karat. The beauty of errors: patterns of error correction in desktop speech systems. In *INTERACT '99*, pages 133–140. IOS, 1999.

[7] S. Hart and C. Wickens. *Manprint: an Approach to Systems Integration*, chapter Workload Assessment and Prediction, pages 257–292. Van Nostrand Reinhold, 1990.

[8] M. Karam and M. C. Schraefel. Investigating user tolerance for errors in vision-enabled gesture-based interactions. In *Proc. AVI '06*, pages 225–232, NY, USA, 2006. ACM.

[9] H. Ketabdar, K. A. Yuksel, A. JahanBekam, M. Roshandel, and D. Skripko. Magisign: User identification/authentication based on 3d around device magnetic signatures. In *Proc. UBICOMM '10*, 2010.

[10] C. Kray, D. Nesbitt, J. Dawson, and M. Rohs. User-defined gestures for connecting mobile phones, public displays, and tabletops. In *Proc. MobileHCI '10*, pages 239–248, NY, USA, 2010. ACM.

[11] J. Mankoff and G. D. Abowd. Error correction techniques for handwriting, speech, and other ambiguous or error prone systems. Technical report, GIT-GVU, 1999.

[12] A. Oulasvirta and J. Bergstrom-Lehtovirta. Ease of juggling: studying the effects of manual multitasking. In *Proc. CHI '11*, pages 3103–3112, NY, USA, 2011. ACM.

[13] S. Oviatt. Taming recognition errors with a multimodal interface. *Commun. ACM*, 43:45–51, 2000.

[14] S. Oviatt, M. Maceachern, and G. anne Levow. Predicting hyperarticulate speech during human-computer error resolution. *Speech Communication*, 24:87–110, 1998.

[15] S. Oviatt and R. VanGent. Error resolution during multimodal human-computer interaction. In *Proc. ICSLP '96*, volume 1, pages 204–207, oct 1996.

[16] J. C. Read, S. MacFarlane, and C. Casey. Oops! silly me! errors in a handwriting recognition-based text entry interface for children. In *NordiCHI'02*, pages 35–40, NY, 2002. ACM.

[17] W. R.G. and N. M. An interface-oriented approach to character recognition based on a dynamic model. *Pattern Recognition*, 31(2):193–203, 1998.

[18] J. Rico and S. Brewster. Usable gestures for mobile interfaces: evaluating social acceptability. In *Proc. CHI '10*, pages 887–896, NY, USA, 2010. ACM.

[19] B. Rime and L. Schiaratura. *Fundamentals of Nonverbal Behavior*, chapter Gesture and speech, pages 239–281. Cambridge University Press, 1991.

[20] J. Ruiz, Y. Li, and E. Lank. User-defined motion gestures for mobile interaction. In *Proc. CHI '11*, pages 197–206, NY, USA, 2011. ACM.

[21] B. Suhm, B. Myers, and A. Waibel. Multimodal error correction for speech user interfaces. *ACM Trans. Comput.-Hum. Interact.*, 8:60–98, 2001.

[22] J. O. Wobbrock, M. R. Morris, and A. D. Wilson. User-defined gestures for surface computing. In *Proc. CHI '09*, pages 1083–1092, NY, USA, 2009. ACM.