Blind FLM: An Enhanced Keystroke-Level Model for Visually Impaired Smartphone Interaction

Shiroq Al-Megren, Wejdan Altamimi, Hend S. Al-Khalifa

Blind FLM: An Enhanced Keystroke-Level Model for Visually Impaired Smartphone Interaction

Shiroq Al-Megren, Wejdan Altamimi, Hend S. Al-Khalifa

Department of Information Technology, King Saud University, Riyadh, Saudi Arabia
{salmegren, hendk}@ksu.edu.sa, wejdanaltamimi@gmail.com

Abstract. The Keystroke-Level Model (KLM) is a predictive model used to numerically predict how long it takes an expert user to accomplish a task. KLM has been successfully used to model conventional interactions, however, it does not thoroughly render smartphone touch interactions or accessible interfaces (e.g. screen readers). On the other hand, the Fingerstroke-level Model (FLM) extends KLM to describe and assess mobile-based game applications, which marks it as a candidate model for predicting smartphone touch interactions.
This paper aims to further extend FLM for visually impaired smartphone users. An initial user study identified basic elements of blind users' interactions that were used to extend FLM; the new model is called "Blind FLM"'. Then an additional user study was conducted to determine the applicability of the new model for describing blind users' touch interactions with a smartphone, and to compute the accuracy of the new model. Blind FLM evaluation showed that it can predict blind users' performance with an average error of 2.36%.

Keywords. Keystroke-level mode (KLM), Fingerstroke-Level Model (FLM), mobile phone, smartphone, mobile KLM, touch interaction, visually impaired users, blind users.

1 Introduction

In Human Computer Interaction (HCI), predictive models allow for human performance to be measured analytically to evaluate the usability of computer systems' design scenarios using low fidelity prototypes and no user participation [19]. The Model Human Processor (MHP) provides a simplified view of the human information processing system that can be used to predict user behaviour. MHP is one of the key components of Card et al.'s [5] framework for human performance modelling, and is a central part of the framework's other key component; a set of techniques collectively referred to as Goals, Operators, Methods, and Selection rules (GOMS). This family of techniques is used to compare and evaluate motor behaviour by describing four components of skilled error-free user performance: goals, operators, methods, and selection rules.

The GOMS family is used to model goal hierarchies of defined unit tasks. The tasks are rendered as a composition of actions and cognitive operations. The analysis of the

composition yields quantitative and/or qualitative measures of performance [5]. Members of the GOMS family differ in their analysis complexity and the accuracy of predicted completion times [11,12]. The Keystroke-Level Model (KLM) is a simplified implementation of GOMS that is used to numerically predict execution times for specific tasks in a desktop environment using mouse and keyboard input [4]. The simple model has been widely applied to predict expert performance of various desktop interfaces and its analysis has proven accurate [23]. This fact demonstrates the aptitude and usefulness of KLM.

Originally intended for desktop systems, KLM has been continually extended to model new paradigms of user interaction. With the advancement of smartphone technologies, the original model has been modified for smartphone interaction to ease and accelerate usability testing in the early phases of development. Smartphone extensions to KLM review the model's decomposition by modifying the original operators, introducing new actions, and revising execution times. These new interactions and extended models include predictive text entry [17], voice recognition [6], Near Field Communication (NFC) technology [10], touch input [20], and touch-less interaction [8].

GOMS techniques and smartphone extensions to KLM model visual desktop and smartphone systems to overcome the drawbacks of usability testing by reducing cost and identifying problems early in the development process. However, these models assume non-disabled users that are able to visually perceive the interface. Visually impaired users [27] utilise assistive tools to decrease or eliminate visual dependency. Screen readers provide auditory descriptions of visual elements on a traditional screen. Similarly, smartphones provide accessible interfaces to compensate for visual impairment. Models formulated exclusively from and for visual computer systems are ill-equipped to represent interfaces accessible to visually impaired users and their interactions.

This paper proposes an extended KLM model that is applicable to visually impaired smartphone interaction, and it makes two main contributions. First, a selected mobile KLM extension is examined and modified to model visually impaired smartphone interactions. Second, the new model is evaluated in two user studies where the model was able to thoroughly render and accurately predict blind users' interactions with a smartphone.

In the following sections, we first describe KLM and recall its benefits and limitations as a predictive model. Next, we review the literature on cognitive models for accessible designs and extensions to KLM. We then present the first study for the purpose of extending KLM. This resulted in an enhancement to FLM that considers accessible designs for blind smartphone users. Furthermore, we validate the extended model in a main experiment. We then present and discuss the results of the validation experiment. Finally, we draw conclusions and future plans.

2 Background and Related Work

The extended KLM for visually impaired smartphone interaction builds upon prior work in predictive models and a stream of KLM expansions in accessible interfaces and

smartphones. While this section does not represent a complete review of the state of the art for model based usability evaluation, it, nevertheless, highlights the most relevant work and how they may differ.

2.1 Keystroke-Level Model (KLM)

KLM is one of GOM's simpler techniques that computes the time it takes an expert user to perform an error-free task on a desktop application. KLM inherits several limitations from GOMS that limit analysis to linear, closed tasks that are executed error-free by expert users. Task execution time is predicted in KLM by decomposing a set of tasks into a list of perceptual, cognitive, or motor operators and computing its summation. The model consists of six operators [4]:

- **Keystroke *K*** key or button press
- **Point *P*** point at a target with the mouse
- **Home *H*** move hands to the home position (keyboard or mouse)
- **Draw *D*** draw a line on a grid
- **Mental act *M*** mental processing prior to taking an action
- **Response *R*** system response time

Moreover, the mental operator is governed by a set of heuristic rules that consider cognitive preparation:

- Rule 0: insert *M* operators in front of all *K* operators. Also, place *M* operators in front of all *P* operators used to select commands.
- Rule 1: remove *M* operators that appear between two operators anticipated to appear next to each other.
- Rule 2: remove *M* operators belonging to one cognitive unit except the first; a cognitive unit is a premeditated chunk of cognitive activities.
- Rule 3: remove *M* operators that precede consecutive terminators.
- Rule 4: remove *M* operators that preceded terminators of commands.

The unit execution time for each operator (excluding *R*) have been set from previous HCI research. KLM predicts a task's execution time by adding the operators' unit times for each of the task's activities, where $T_{operator}$ is an operator's total time:

$$T_{execute} = T_K + T_P + T_H + T_D + T_M + T_R$$

KLM was empirically validated against keyboard and mouse based systems and various tasks [4]. The model's predictions were found to be accurate with an error of approximately 21%.

2.2 Cognitive Models in Accessibility

The design of accessible web pages are governed by sets of regulations and guidelines to maximise its use among users of varying capabilities [26]. Automated tools are utilised by designers to assess compliance, however these tools only evaluate checkpoints and do not thoroughly assess other usability issues (e.g. effectiveness and efficiency). Visually impaired computer users use screen readers to navigate applications; visual content is represented as a coded linear sequence that is synthesised into auditory presentation. The GOMS family of techniques are suited for modelling screen readers' sequential output as the model's application is limited to linear tasks. A handful of research have explored extending predictive models to reproduce screen readers' auditory representation.

Time-oriented aspects of usability were the focus of a new visualisation approach, Blind Usability Visualisation [22]. This approach was later implemented as a disability simulation tool, Accessibility Designer (aDesigner), to evaluate and visualise the usability of web pages for blind users. The tool's most novel feature is the concept of 'reaching time'; the time it takes a blind user using a screen reader to reach a desired destination on a web page from the top of that page. The tool is not a predictive model and while average reaching time can be used to measure the navigability of a web page by blind users, it excludes cognitive decision times or other operations. An extension to one of GOMS techniques addresses this rough estimation of reaching time [24].

User observations and two field studies were performed to provide a broad overview of blind interaction on accessible web pages [24]. The studies identified key findings of blind interactions, this include: reliance on different navigational strategies, frequent speech rate configuration, verification of screen reader output, and activation of interactive elements. Some of these findings were used to introduce new structures to the Natural GOMS Language (NGOMSL) that extended the model for accessible web pages. Configurable speech rates and Braille readings times and their impact were not considered. This modified model aimed to automate the assessment of accessible web page efficiency by calculating the time it takes to execute a task on a web page. Nevertheless, the model remains qualitative in nature and unverifiable which makes its application difficult.

Working with non-disabled users, new models were introduced to assess mouse and keyboard navigation in a web site [21]. The keyboard model focused on users who could not use a pointing devices and quantified keyboard navigation's disadvantage against mouse navigation. The TAB key is used to navigate a chain of links in a web page by first locating the target link then pressing the key n times until the link is reached and finally activated by pressing ENTER. This is clearly problematic for link-intensive web sites. The models extend KLM, each of which introduced new operators. Time estimates for the new keyboard operators were measured in a laboratory experiment with non-disabled expert users. Theoretically the model can be used to render blind interaction, but requires adaptation to consider the time it takes a blind user to hit the TAB key.

Blind users' interaction with web pages via a screen reader were remotely observed and analysed to supplement KLM [25]. CogTool [13], a cognitive modelling tool that

supports rapid evaluation analysis of GOMS formulations, was used to validate KLM's efficacy at modelling blind users' interaction. KLM did not accurately model the user's behaviour as it did not consider the screen reader's high speech rate. Additionally, CogTool's rules for placing KLM's mental operators, which were procured from sighted user's interaction with visual content, did not ideally reflect the observed skilled interaction of the blind user. Discarding the mental operators from calculation did not improve accuracy, which suggests the parallel recognition of situations and decisions, as well as auditory reception of the screen reader's speech. A future tool was envisioned for evaluating blind users' interaction on a web page, but was not implemented.

2.3 KLM Extensions

Usability testing is an expensive process that is exacerbated with disabled users as testing will have to be carried out late in production with high fidelity prototypes. The recent direction of model extension address the usability needs of non-disabled smartphone users. These models lend themselves to further modification to address blind smartphone interaction. KLM is typically extended by evaluating original operators and introducing new operators and equations. This section focuses on smartphone extensions to KLM and reviews direct touch models.

In the context of smartphone interaction, studies for extension began with text entry methods and predictive text. An early model extended KLM for three text entry methods using a smartphone's keyboard and compared predictions of typing speeds for each of these methods [7]. Another model identified and validated new operators that represent typical and advanced smartphone interaction (e.g. identification tags and gestures) [9]. The model was later revised to include NFC interactions [10]. An extended model utilised KLM, Fitts' law, and a language model to predict user performance with two types of Chinese input methods on smartphones [17]. For the purpose of presenting a new keyboard, 1Line, KLM was extended to measure multi-finger touchscreen keystrokes [15]. Replacing the keyboard with speech input, a new model investigated the feasibility of a speech-based smartphone interface for text messaging, which adapted original operators and introduced predictive equations to the model [6].

Beyond keyboard or speech input, KLM had been extended to predict user interaction time and system energy consumption on smartphones [18]. KLM was also adapted for next-generation smartphone designs, particularly phones that utilise styli [16]. The new model introduced new operators that uniquely represent stylus interactions, and presented the concept of operator block (a sequence of operators that can be used with high repeatability). For direct touch interaction, KLM was modified to model middle-sized touch screens in Integrated Control Systems (ICSs) [1], where the prediction error was less than 5%.

Touch-based smartphones later replaced traditional phones where new extensions were required to assess this new paradigm of human interaction. The Touch Level Model (TLM) was proposed to support interaction with touch devices via direct interaction [20]. Several operators were retained from KLM as they remain applicable to touch input (keystroking *K*, homing *H*, mental act *M*, and system response time *R*), but discarded the drawing *D* operator. Several new operators were freshly introduced or

inherited from other extensions to KLM that were not developed for touch input: distraction *X* [9], gesture *G*, pinch *P*, zoom *Z*, initial act *I* [9], tap *T*, swipe *S*, tilt *L(degrees)*, rotate *O(degrees)*, and drag *D*. TLM has the potential for benchmarking users' touch interactions, but the new operators are without baseline values and the model has yet to be validated. Retained operators' unit times will likely need to be reexamined as well.

Fitts' law is a descriptive model that considers the physical aspects of a Graphical User Interface (GUI) and predicts the time it takes to point to a target. In previous research, Fitts' law has been integrated with KLM to produce enhanced smartphone versions of the model and to compute average execution time based on physical interface features (e.g. [17]). One enhancement extended KLM with three common touch interactions: swipe, tap, and zoom [3]; interactions similarly modelled in TLM [20]. Unlike TLM [20], unit operators' times were formulated using Fitts' law. Nevertheless, the model's potential has not been verified against their intended interaction and application.

Mobile games have increased in popularity as smartphones are becoming more durable and supportive of direct touch interaction. The Fingerstroke-Level Model (FLM) is a modified version of KLM developed for the evaluation of mobile gaming efficacy [14]. FLM adapted original operators and introduced new ones to cope with the new interactions. The model is comprised of six operators: tap *T*, point *P*, drag *D*, flick *F*, mental thinking *M*, and response time *R*. FLM shares *P*, *M*, and *R* with the original KLM, and tapping and dragging (i.e. swiping) with TLM [20] and El Batran et al.'s extension [3]. Unlike the original KLM and its extensions [3,20] that results in a single deterministic value, FLM is a regression model. FLM was applied to a mobile game where it was able to predict its execution time more accurately than KLM.

3 KLM Extension for Blind Interaction

The main objective of the first user study was to explore blind users' interaction with touch-based smartphones. Prior to the study, an online screening questionnaire was distributed to better understand blind users' smartphone interaction. The questionnaire's main objectives were to identify commonly used smartphones, popular applications, and blind users' experience with smartphones. The questionnaire was conducted in Arabic and garnered twenty-one respondents, the majority of which were female with an average age of 26.57 years (standard deviation, SD = ±9.66). Excluding one participant, the entire sample used iPhones with the majority (81%) having at least three years of experience with the device. The respondents ranked Twitter, WhatsApp, and YouTube as their most often used applications.

3.1 Methodology

Two instruments were used in this study: structured interviews and observation. The interview was designed to discover popular actions utilised by blind users when using applications on a smartphone. The vocalised actions were then confirmed via observation.

Participants. Three female blind participants with a mean age of 20 years took part. The participants were university students in the College of Education at King Saud University. All participants had good experience with using an iPhone (average experience of 5.6 years) and gave vocal informed consent.

Apparatus. The iPhone was screened as the most commonly used smartphone and its use was observed in this study. VoiceOver is a built-in speech synthesiser that assists visually impaired users when interacting with iOS devices (e.g. iPhone) and applications. Along with specific gestures, users are able to navigate and activate interface element. For instance, a user taps to select an element and listen to its auditory description which informs upcoming actions. Keyboard input is also facilitated with audio. Speech rate of the synthesiser is adjustable in VoiceOver via a rotor that is manipulated by rotating two fingers on the screen. However, this action is often infrequent during a task and is therefore excluded from consideration.

Materials. Interview questions were predetermined starting with an introductory question regarding the smartphone used and its version. The following question prompted the participant for any set of actions that are typically adopted with iPhone applications. The participant was asked to list these actions if she was able to identify a particular set of actions repeatedly used among various application. Otherwise, the participant is asked to describe her interaction with iPhone applications. For the study's instrument, two tasks for Twitter and WhatsApp were prepared for observation. In the Twitter task, the participant was asked write and send a tweet on her personal account. For WhatsApp's task, the participant was asked to write and send a message to someone from her contact list.

Procedure. Sessions were held in a quiet room and lasted approximately 40 minutes. First, the participant was welcomed and the general research idea was introduced. Two instruments were used in this study, interviews and observations, which were conducted in the same session in sequence. First, the questions were put forth to the participant in Arabic. Second, the participant was asked to carry out the Twitter or WhatsApp scenario under observation. The choice of task and application was dependent on the participant's familiarity with said applications.

3.2 Results

From the interviews, all three participants agreed that they did not follow a series set of actions when interacting with their iPhones. Nevertheless, when asked to describe the sequence of actions typically taken when interacting with an application, the participants identified the following sequence: listen, navigate to a certain button or content (via flick operation), then activate the element (via double tap). These reported actions were then verified with observation.

Two of the participants performed the WhatsApp task, while the third participant carried out the Twitter task. In the Twitter task, the participant opened the application with a double tap and then navigated within the application by flicking the screen with her finger. At times, the participant listened to the complete audio description of the visual element before deciding on an action. Other times, the participant was satisfied with a partial description. For text input, the participant tapped on the screen until the textfield was located (this was vocalised with VoiceOver) and used double tap to activate the keyboard. The writing process started with a tap on the approximate position of the intended character to hear the description. These actions were similarly observed with the WhatsApp task.

Blind users' smartphone interaction via the device's screen reader can be summarised into four actions: tap, double tap, flick, and drag. Tap actions are used to select an element. The selected element is activated via a double tap action. Flicking a finger on the screen is used to navigate the application's elements. Vertical and horizontal scrolling is achieved by sliding or dragging three finger across the screen.

3.3 Revised FLM for Blind Users

The user study identified a series of operations that were frequently carried out by blind users interacting with a smartphone application: tap, double tap, flicking a finger on the screen, or scrolling vertically or horizontally by dragging three fingers across the display. These actions were mapped against the previously reviewed touch-based smartphone extensions to KLM (see Section 2.3 and Table 1).

Table 1. The observed blind actions mapped against the same/similar actions (i.e. operators) in TLM [20], El Batran et al.'s model [3], and FLM [14].

Action/Model	TLM [20]	El Batran et al. [3]	FLM [14]
Tap	Tap *T*	Tap	Tap *T*
Flick	Swipe *S*	Short swipe	Flick *F*
Double tap			
Drag	Drag *D*	Long swipe	Drag *D*

The four observed actions related to three operators from TLM [20], El Batran et al. [3], and FLM [14]. Tap actions are a common direct touch behaviour that corresponds well to the Keystroke *K* operator in KLM and was identified in the previous literature [3,14,20]. The observed flick action was a single finger swipe that is short and quick and directly mapped against FLM's Flick *F* [14]. TLM [20] defined Swipe *S* as placing one or more fingers on the screen and moving that finger in a single direction for a period of time, while El Batran et al. [3] described swipe as a short or long action to achieve tasks such as scrolling. These two actions can roughly represent the observed flick action, while TLM's [20] Drag *D* and El Batran et al.'s [3] long swipe can model the observed drag action.

Of the three touch-based smartphone extension to KLM, FLM's [14] operators closely resemble the observed actions. Unlike TLM [20], units times were computed for the various operators in FLM [14] and El Batran et al. [3]. However, El Batran et al.'s [3] model only provides unit time for a short swipe (resembling a flick) and not for a long swipe. Moreover, the operators in FLM were validated in an experiment where the root mean square error (RMSE) of the observed and predicted execution times was 16.05%. For that purpose, FLM was selected as the prime candidate for extension. The extended FLM model is called Blind FLM, and its operators and unit times are summarised in Table 2 in relation to KLM [4] and FLM [14].

Table 2. Retained, excluded, and new operators of the extended FLM (Blind FLM) and their time estimates as compared to the original KLM [4] and FLM [14]. [1] Drawing *D* assumes *n* straight line segments having a total length of *l*. [2] Flick *F* value is set at 0.12 seconds considering error-free navigation that were observed to be typically from right to the left.

KLM [4]	Time (s)	FLM [14]	Time (s)	Blind FLM	Time (s)
Keystroke *K*	0.2	Tap *T*	0.31	•	0.31
Point *P*	1.1	•	0.43		
Draw *D*	$0.9n + 0.16l$[1]	Drag *D*	0.17	•	0.17
Home *H*	0.4				
Mental act *M*	1.35	•	1.35	•	1.35
Response *R*	variable	•	variable	•	variable
Extensions					
		Flick *F*	$0.12_{\text{right-to-left}}$ $0.11_{\text{left-to-right}}$	•	0.12[2]
				Double tap *DT*	0.62

Retained Operators. Five of the original FLM operators are still appropriate to model blind users' interactions on an iPhone mobile device with VoiceOver.

- **Tap *T*** A blind user taps anywhere on the smartphone's screen to listen to the audio description of the underlying visual element. This could be a button, text, link, image, or video. Tap is also used to choose the start position for navigation.
- **Drag *D*** Vertical and horizontal scrolling is performed by sliding/dragging three fingers across the screen.
- **Flick *F*** Unlike dragging *D*, this action is typically quick and achieved with a single finger to navigate application elements.
- **Mental preparation *M*** With the absence of sight, blind users rely on other senses to conceptualise the real world. Blind smartphone users utilise audio description to map their next interaction, i.e. the mental preparation needed to perform the following action. KLM [4] was previously refined to model screen readers and the authors suggested that recognition of the present situation, screen reader's speech, and action decisions occur in parallel. This was also argued for typing actions [25].

- **Response time *R*** This operator is system dependent, arguably irrelevant due to the technological advancement and negligible response times. Nevertheless, the variable is still retained to account for different devices and software.

New Operators. New operators are introduced to the extended FLM to account for novel interactions that are afforded by the analysed interface. Tap interactions are frequently utilised by blind users to select an element and is retained from FLM [14]. Double tap actions activate the selected element (via tap). This is unique to blind users' interaction, where the former voices the element and the latter launches the element.

Excluded Operators. The pointing operator *P* is excluded in Blind FLM as it is not applicable to blind users' interaction since pointing at an element requires visual perception. Instead of pointing, a blind user taps close to a target element or navigates the elements sequentially. Both of which are represented by the original FLM operators: tap and flick, respectively [14]. Thus, the act of pointing works in tandem or is encapsulated with/within the subsequent action and is not used in its singularity.

Operators' Unit Times. The baseline value for the FLM [14] operators were computed with a practical study. The original values are employed for Blind FLM. In the case of the double tap *DT* action, the value of tap *T* is multiplied by two. The mental thinking *M* operator was not changed from the original KLM and is maintained for this extension as well. See Table 2 for execution times for all operators.

4 Blind FLM Validation

The efficacy of a model and the accuracy of its baseline values are typically evaluated through controlled research studies with human subjects. Within-participants experiments were conducted to investigate Blind FLM. The purpose of these two studies were twofold: 1) to determine the applicability of the Blind FLM operators for describing blind users' touch interactions with a smartphone (first user study); 2) to compute the accuracy of the new model by comparing observed execution times with predicted times (second user study).

4.1 First User Study

A preliminary study was carried out to satisfy the first purpose of the experiment; determine if the new model and its operators are fit to fully model blind touch interactions with a smartphone. This study was also used to evaluate the experimental tasks in order to refine the tasks for the next study.

Methodology.

Participants. Five female participants with a mean age of 20.2 years (SD = ±0.84) were recruited for the experiment. All participants were familiar with using an iPhone and VoiceOver with an average of 5.8 years of experience (SD = ±1.1). The participants were students recruited from King Saud University. The device's VoiceOver speech rate values ranged from 80% to 100% with an average of 90%.

Apparatus. Apple's iPhone 6 with VoiceOver was used by all participants. Access to the three applications were made via the participants' private accounts. A camera was used to video record participants' interactions.

Task. Based on the previously discussed questionnaire results: Twitter, WhatsApp, and YouTube were the top three used applications in the blind community. For the preliminary study, three sessions were dedicated for each of these applications. Each session consisted of three tasks; one open task and two structured tasks (a total of nine tasks).

- Twitter
  - Structured tasks
    1. View the profile of the first account on the 'Following' page
    2. Write a tweet consisting of a single word in Arabic or English (e.g. 'Hello') and to tweet the message
  - Open task: retweet any tweet from the timeline
- WhatsApp
  - Structured tasks
    1. Make a voice call with the first contact from the chat list
    2. Reply to the first chat from the chat list with a single word in Arabic or English (e.g. 'Hello')
  - Open task: create a new chat group with two contacts
- YouTube
  - Structured tasks
    1. Play the first video in the home page
    2. Subscribe to the channel of the first video in the home page
  - Open task: delete YouTube's browsing history

The open scenario was used to determine if Blind FLM was able to thoroughly represent blind users' interaction with a touch-based smartphone. KLM and extensions of KLM are only equipped to represent error-free interactions and tasks, thus the two structured tasks were used to reduce the space of probability and error. The two structured tasks were presented as a set of steps that begin from a uniform starting point that continued sequentially.

Procedure. Sessions were held in a quiet room with a WiFi connection, and each session lasted approximately 45 minutes. First, the participant was given verbal instructions about the experiment and its purpose. After which, the participant was given a

chance to voice any questions about the study. Consent was then collected and recorded verbally. Each of the nine tasks was presented to the participant with its required steps. Prior to starting a task, the participant practiced the task until it was mastered and completed without error. For the structured task, whenever a mistake occurred the participant was asked to redo the scenario. All sessions were video recorded.

Participants were asked to use their own iPhones for the experiment. To ensure uniform VoiceOver settings for all participants a set of instructions were provided. The speech rate of VoiceOver was set on 80% speed. The volume of the speech synthesiser was set to its highest rate. Typing mode was set to standard typing.

Results. The video recording for all sessions were coded using the Behavioural Observation Research Interactive Software (BORIS) [28]. BORIS, an event logging tool, allowed the experimenter to observe the video recordings and log observations, i.e. Blind FLM's operators.

Structured Tasks. The structured tasks were modelled using Blind FLM to predict execution times. All tasks were first modelled with the proposed Blind FLM physical/motor operators. The response time *R* operator was not used due to the iPhone's almost instantaneous reaction. Mental act *M* operators were later added based on the new model's modified heuristic rules. Rules 1, 3, and 4 from the original KLM are not applicable in this context. Rule 1 is related to fully anticipated operators, while Rules 3 and 4 handle syntactic terminators. Rule 0 and 2 were modified from the original KLM to reflect visually impaired smartphone interaction:

- Rule 0 (R0 base rule): insert *M* operators in front of all *K* operators that are used to type a text. Also, place *M* operators in front all *P* operators that are used to select a method. In Blind FLM, flick *F*, tap *T*, double tap *DT*, and drag *D* operators substitute keystroke *K* and point *P* operations.
- Rule 2 (R2): remove all *M* operators that are related to one cognitive unit except the first *M*.

For Twitter's first task, one participant was excluded from the analysis due to multiple mistakes. The observed execution time for that scenario with the four participants was 13.8 seconds compared to a predicted value of 11.33 seconds. For the first WhatsApp structured task, the average predicted execution time was 7.72 seconds compared to an observed value of 8.34 seconds. The first YouTube task's predicted execution time was 4.04 seconds compared to the observed value of 5.68 seconds. The second tasks from all three applications were excluded due to varying typing speed in Twitter and WhatsApp, and for repeated mistakes in YouTube's task. The root mean square error (RMSE) is commonly used to evaluate KLM's predictions [4]. The average computed RMSE for the three structured tasks was 1.73%.

Open Tasks. The average number of actions undertaken for the open task was 14.6 for Twitter, and 19 for YouTube, and 60.8 for WhatsApp. Negligible mistakes were detected, with no more than two errors per task. The Twitter task consisted of 73% flick

F actions, 19% double taps *DT*, and 8% tap *T* actions. For the WhatsApp scenario, approximately 40% of the interactions were categorised as flick *F* and 30% were tap *T* actions. Double tap *DT* made up 26% of the actions, while drag actions *D* were only performed 4% of the time. The majority of the interactions in the YouTube task were flick *F* actions (78%), followed by double tap *DT* (21%) then tap *T* (1%). No drag *D* actions were observed for the Twitter or YouTube tasks. The open tasks were not analysed with Blind FLM as participants were not restricted to a set of predefined task sequences.

Discussion. Blind FLM was well-equipped to model blind user's touch interactions with an iPhone and VoiceOver. This section discusses the findings of the preliminary study for the open and structured tasks.

Structured Tasks. In the original KLM, the observed model error was 21% of the average predicted execution time [4]. This level of accuracy was achieved in the preliminary study. Nevertheless, the sample size was too small to be representative. Additionally, it was observed with the structured tasks that typing speed varied between the participants which may affect the consistency of the computed results. The second structured task for each of the examined applications involved a typing subtask. These tasks will be excluded in the upcoming user study. Unlike the other actions, modelling text entry is complex. Previous findings for text entry identified various factors that impact text entry, this includes: repetition effect (first tap, second, or more), key type (number, alphabet, or character), entry method (e.g. predictive or word completion), typing speed, and language corpus. Due to this distinctiveness, text entry is excluded with a plan for a future extension.

Open Tasks. The open task for each of the three applications were observed to determine Blind FLM's coverage of blind users' interactions. The operators' occurrence rates were computed and indicated the operators' priorities and their weights in the total execution time. The majority of operators were regularly used to model interactions. Of those operators, flick *F* was the most frequently used when interacting with the smartphone. The drag *D* action was the least used operator and was only observed in WhatsApp's open task.

4.2 Second User Study

A second user study was carried out to validate Blind FLM. The purpose of this study was to compute the accuracy of the new model by comparing observed execution times with predicted times. The experiment expanded on parts of the previous study and improved the tasks to overcome inaccuracy concerns.

Methodology.

Participants. Twenty right-handed individuals (7 males, 13 females) with a mean age of 21.5 years (SD = ±7.24) took part in the experiment. Participants were recruited from Kafeef organisation (an organisation that is concerned with training and qualifying blind citizens), Alnoor institute (a female school for the blind), and King Saud University. On average the participants had 5.4 years of experience (SD = ±1.5) with using an iPhone and VoiceOver. The average speed rate utilised with VoiceOver was 89%.

Apparatus. Similar to the previous study, an iPhone 6 was used by all participants with VoiceOver. Each participant used their own device, as well as accessed their private Twitter, WhatsApp, and YouTube accounts. The sessions were video recorded with a camera.

Task. The task used with this study were previously examined in the previous experiment. For each of the three applications, a single structured task was used. The steps used in the preliminary study were re-evaluated as new versions of WhatsApp and YouTube were released. The starting position for each of the three tasks was the default position of VoiceOver. In Twitter, the participant was expected to follow the steps necessary to display the profile of the first account on the 'Following' page. The task's steps for Twitter were as follows:

- Flick to the settings button
- Flick to the switch account button
- Flick to edit profile button
- Flick to profile name button
- Flick to user account button
- Flick to location button
- Flick to following button
- Double tap on the following button
- Tap on the screen
- Scroll down once
- Tap on the screen
- Double tap on the first account

For the WhatsApp task, the participant was asked to initiate a video call with the first contact on their chat list (see Table 3 for steps). For the YouTube session, the participant was asked to play the first video on their personal home page by following these steps:

- Flick to 'Upload and Record'
- Flick to the search button
- Flick to personal account
- Flick to first video on the home page
- Double tap on video to play

Table 3. WhatsApp's task modelled with Blind FLM.

Task steps	Operator	Predicted unit time (s)
	M	1.35
Flick to chats	*F*	0.12
Flick to the compose button	*F*	0.12
Flick to archived chats	*F*	0.12
Flick to search field button	*F*	0.12
Flick to broadcast list	*F*	0.12
Flick to new group button	*F*	0.12
Flick to first chat	*F*	0.12
	M	1.35
Double tap on chat to open	*DT*	0.62
	M	1.35
Flick to the video call button	*F*	0.12
	M	1.35
Double tap on button to call	*DT*	0.62
	Predicted execution time	7.6

Procedure. Sessions were held in a quiet room with a WiFi connections, the location varied depending on the participants and their affiliation. Each session lasted approximately 20 minutes. The participant was welcomed and the purpose of the study was explained. The participant was then given a chance to voice any questions. Next, a consent form was vocalised and the participant's response recorded. Each task was presented to the participant and its steps were explained. The participant was asked to train each task until mistakes are no longer made. The participant was also asked to use his/her own iPhone for the study with a VoiceOver speech rate set on 80%. The speech synthesiser volume was set to the highest rate. All sessions were video recorded.

Results. The three tasks were modelled with Blind FLM to predict execution times (see sample model for WhatsApp's task in Table 3). The tasks were first modelled without any mental thinking *M* operators. The mental thinking *M* operator was then incorporated in the model following the modified heuristic rules for *M* insertion. Observed execution times were logged using BORIS, where video recordings were played frame by frame. Twitter's task was predicted to take 10.97 seconds and its execution time was observed at 13.28 seconds. The predicted execution time for the WhatsApp task was 7.6 seconds (see Table 3) compared to the observed value of 8.7 seconds. For YouTube's task, execution time was predicted at 3.8 seconds, while the observed value was 4.8 seconds. The accuracy of Blind FLM was evaluated using RMSE, which showed an average prediction error of 2.36%. Table 4 shows the predicted and observed execution times for the three tasks, as well as the RMSE for each of these tasks and the average RMSE.

Table 4. Average observed time, predicted execution time, and computed RMSE for each of the three experimental tasks, as well as average RMSE.

Task	Observed time (s)	Predicted time (s)	RMSE
Twitter	13.28	10.97	3.81%
WhatsApp	8.7	7.6	2.01%
YouTube	4.8	3.8	1.27%
		Average RMSE	2.36%

Discussion. The results show that all three tasks completion times were under KLM's suggested 21% RMSE [4]. For all tasks combined, the RMSE average was 2.36%. Twitter's task was predicted and observed to be the longest, while YouTube's task took the least time. This was also reflected with the RMSE value where a larger error percentage was incurred for Twitter's task. This indicates the effect of task complexity on the model's prediction error percentage. See Table 4.

The mental thinking *M* operator's unit value used in Blind FLM is the original value assigned to it with KLM [4]. Many studies similarly retained the original unit measure for M for conventional phones [7] and smartphones [14,20]. However, these models render visual applications. A previous study observed blind users' web page interactions via a screen reader to extend KLM [25]. The results of the study suggested the parallel recognition, decision, and auditory reception of the screen readers speech, directly affecting the unit time of the model's operators. While this was also observed in the study (e.g. a participant not listening to the complete description of a visual element), completely discarding the mental act *M* operator did not result in better accuracy. This added complexity to the mental act *M* operator for accessible interfaces merits a revisit in future research.

For all tasks, Blind FLM underestimated the predicted time (see Table 4). Given that KLM can only model error-free tasks (a limitation passed on to its extensions, including Blind FLM), participants were asked to repeat a task until no errors were recorded. This repetition had lead the participants to caution and it indicates that the observed times are likely upper limits. Thus, the observed extra time highlight the time it took the participants to process the instructions and carry out the task, i.e. contrary to natural interaction.

5 Conclusion and Future Work

KLM and mobile extensions to KLM have popularly been adopted in HCI to predict the time it takes a skilled user to complete an error-free task. Skilled users were inherently assumed to be sighted and the applications relied on visual output. These models cannot thoroughly express visually impaired interaction, where the visual interface is replaced with an audible alternative. This paper makes two contributions. First, FLM [14] was modified further to render visually impaired mobile phone interaction. Some of FLM's operators were retained, while other were excluded. A new operator, double tap *DT*, significant to blind user's interaction was introduced to the model. These

changes were applied after a series of interviews and observations with blind users and their touch-based smartphones. Second, a user study was conducted to validate the new model (Blind FLM), and evaluate its predictions on a number of tasks. The results showed that Blind FLM was able to accurately predict execution time well below the suggested 21% error [4].

The user studies carried out to modify KLM and validate the new model were performed with an iPhone and limited to a set of applications (WhatsApp, YouTube, and Twitter). We intend to evaluate Blind FLM's accuracy with other mobile phones and applications. The model's operators' unit times were not measured and instead original values from KLM [4] and FLM [14] were adopted. In future work, we plan to reexamine these values in controlled human-subject trails. This is particularly important to overcome any inaccuracies due to FLM and Blind FLM's varying application domains. It will also be interesting to reevaluate the mental thinking *M* operators' unit time and the effects of an audible interface and blind users listening abilities on its value [2]. Due to varying typing speeds, typing tasks in the validation studies were excluded from computation. This exclusion lends itself to future research that considers visually impaired text input and its effect on tap *T* and double tap *DT* operators. A final future direction will apply Blind FLM on real-life design cases, where the model is used to compare the designs efficacy to make informed design choices.

6 References

1. Evgeniy Abdulin. 2011. Using the keystroke-level model for designing user interface on middle-sized touch screens. *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems*: 673–686. https://doi.org/10.1145/1979742.1979667
2. Chieko Asakawa, Hironobu Takagi, Shuichi Ino, and Tohru Ifukube. 2003. Maximum Listening Speeds for the blind. In *Proceedings of the International Community for Auditory Display*, 276–279.
3. Karim El Batran and Mark D Dunlop. 2014. Enhancing KLM (Keystroke-level Model) to Fit Touch Screen Mobile Devices. In *Proceedings of the 16th International Conference on Human-computer Interaction with Mobile Devices & Services* (MobileHCI '14), 283–286. https://doi.org/10.1145/2628363.2628385
4. Stuart K Card, Thomas P Moran, and Allen Newell. 1980. The Keystroke-level Model for User Performance Time with Interactive Systems. *Commun. ACM* 23, 7: 396–410. https://doi.org/10.1145/358886.358895
5. Stuart K Card, Allen Newell, and Thomas P Moran. 1983. *The Psychology of Human-Computer Interaction*. L. Erlbaum Associates Inc., Hillsdale, NJ, USA.
6. Anna L Cox, Paul A Cairns, Alison Walton, and Sasha Lee. 2008. Tlk or txt? Using voice input for SMS composition. *Personal and Ubiquitous Computing* 12, 8: 567–588. https://doi.org/10.1007/s00779-007-0178-8
7. Mark D Dunlop and Andrew Crossan. 2000. Predictive text entry methods for

mobile phones. *Personal Technologies* 4, 2: 134–143. https://doi.org/10.1007/BF01324120

8. Orlando Erazo and José A Pino. 2015. Predicting Task Execution Time on Natural User Interfaces Based on Touchless Hand Gestures. In *Proceedings of the 20th International Conference on Intelligent User Interfaces* (IUI '15), 97–109. https://doi.org/10.1145/2678025.2701394
9. Paul Holleis, Friederike Otto, Heinrich Hussmann, and Albrecht Schmidt. 2007. Keystroke-level Model for Advanced Mobile Phone Interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '07), 1505–1514. https://doi.org/10.1145/1240624.1240851
10. Paul Holleis, Maximilian Scherr, and Gregor Broll. 2011. A Revised Mobile KLM for Interaction with Multiple NFC-tags. In *Proceedings of the 13th IFIP TC 13 International Conference on Human-computer Interaction - Volume Part IV* (INTERACT'11), 204–221. Retrieved from http://dl.acm.org/citation.cfm?id=2042283.2042306
11. Bonnie E John and David E Kieras. 1996. The GOMS Family of User Interface Analysis Techniques: Comparison and Contrast. *ACM Trans. Comput.-Hum. Interact.* 3, 4: 320–351. https://doi.org/10.1145/235833.236054
12. Bonnie E John and David E Kieras. 1996. Using GOMS for User Interface Design and Evaluation: Which Technique? *ACM Trans. Comput.-Hum. Interact.* 3, 4: 287–319. https://doi.org/10.1145/235833.236050
13. Bonnie E John, Konstantine Prevas, Dario D Salvucci, and Ken Koedinger. 2004. Predictive Human Performance Modeling Made Easy. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '04), 455–462. https://doi.org/10.1145/985692.985750
14. Ahreum Lee, Kiburm Song, Hokyoung Blake Ryu, Jieun Kim, and Gyuhyun Kwon. 2015. Fingerstroke time estimates for touchscreen-based mobile gaming interaction. *Human Movement Science* 44: 211–224. https://doi.org/http://dx.doi.org/10.1016/j.humov.2015.09.003
15. Frank Chun Yat Li, Richard T Guy, Koji Yatani, and Khai N Truong. 2011. The 1Line Keyboard: A QWERTY Layout in a Single Line. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology* (UIST '11), 461–470. https://doi.org/10.1145/2047196.2047257
16. Hui Li, Ying Liu, Jun Liu, Xia Wang, Yujiang Li, and Pei-Luen Patrick Rau. 2010. Extended KLM for Mobile Phone Interaction: A User Study Result. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems* (CHI EA '10), 3517–3522. https://doi.org/10.1145/1753846.1754011
17. Ying Liu and Kari Jouko Räihä. 2010. Predicting Chinese Text Entry Speeds on Mobile Phones. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '10), 2183–2192. https://doi.org/10.1145/1753326.1753657
18. Lu Luo and Daniel P Siewiorek. 2007. KLEM: A Method for Predicting User Interaction Time and System Energy Consumption During Application Design. In *Proceedings of the 2007 11th IEEE International Symposium on Wearable Computers* (ISWC '07), 1–8. https://doi.org/10.1109/ISWC.2007.4373782

19. I Scott MacKenzie. 2003. Motor behaviour models for human-computer interaction. *HCI models, theories, and frameworks: Toward a multidisciplinary science*: 27–54.
20. Andrew D Rice and Jonathan W Lartigue. 2014. Touch-level Model (TLM): Evolving KLM-GOMS for Touchscreen and Mobile Devices. In *Proceedings of the 2014 ACM Southeast Regional Conference* (ACM SE '14), 53:1--53:6. https://doi.org/10.1145/2638404.2638532
21. Martin Schrepp and Patrick Fischer. 2006. A GOMS Model for Keyboard Navigation in Web Pages and Web Applications. In *Computers Helping People with Special Needs: 10th International Conference, ICCHP 2006, Linz, Austria, July 11-13, 2006. Proceedings*, Klaus Miesenberger, Joachim Klaus, Wolfgang L Zagler and Arthur I Karshmer (eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 287–294.
22. Hironobu Takagi, Chieko Asakawa, Kentarou Fukuda, and Junji Maeda. 2004. Accessibility Designer: Visualizing Usability for the Blind. In *Proceedings of the 6th International ACM SIGACCESS Conference on Computers and Accessibility* (Assets '04), 177–184. https://doi.org/10.1145/1028630.1028662
23. Leonghwee Teo and Bonnie E John. 2006. Comparisons of Keystroke-level Model Predictions to Observed Data. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems* (CHI EA '06), 1421–1426. https://doi.org/10.1145/1125451.1125713
24. Henrik Tonn-Eichstädt. 2006. Measuring Website Usability for Visually Impaired People-a Modified GOMS Analysis. In *Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility* (Assets '06), 55–62. https://doi.org/10.1145/1168987.1168998
25. Shari Trewin, Bonnie E John, John Richards, Cal Swart, Jonathan Brezin, Rachel Bellamy, and John Thomas. 2010. Towards a Tool for Keystroke Level Modeling of Skilled Screen Reading. In *Proceedings of the 12th International ACM SIGACCESS Conference on Computers and Accessibility* (ASSETS '10), 27–34. https://doi.org/10.1145/1878803.1878811
26. Web Accessibility Initiative (WAI). Introduction to Web Accessibility. Retrieved January 21, 2017 from https://www.w3.org/WAI/intro/accessibility.php
27. World Health Organization (WHO). Visual impairement and blindness.
28. Behavioral Observation Research Interactive Software (BORIS). Retrieved January 8, 2017 from http://www.boris.unito.it/