# Multimodal Interfaces for Mirco User Actions

Eran Aharonson – Intuitive User Interfaces Ltd. Israel and Department of Software Engineering, Afeka, Tel Aviv Academic College of Engineering

Vered Aharonson – Afeka Center for Language Processing, Afeka, Tel Aviv Academic College of Engineering.

Relevant categories: Multimodal Interaction and Human-Machine Interaction

## Introduction

As users in the 21st century learned to demand not only functionality and performance but also agreeable user experience, the design of consumers' technology is becoming more and more concerned with this aspect. User interfaces meet an even bigger challenge when mobile technology is considered: many interfaces that are acceptable in an office or home environment, are intolerable or even un-usable while on the move and with small devices.

One of the main concepts employed to tackle the interaction issues in small and/or mobile devices is multimodality. Mobile devices today provides various methods of Interaction such as key press, touch, voice control, gestures, tilt control and more. Multimodal interfaces have been shown to have many advantages [1]: they bring robustness to the interface, help the user to correct errors or re-cover from them more easily, and add alternative communication methods to different situations and environments.

Previous studies of human-computer interaction, in both academia and industry, describe research on multi-modal usability issues, but the challenges, although pointed out, are far from being solved. Due to the complexity of the issue, different aspects were studied separately. Those aspects varied from users' reaction to system errors (i.e. [2],[3]) to environmental and social issues of mobility (i.e. [4]). The analysis in the studies varies from qualitative ([1]) to mathematical ([2-6]). The number of parameters involved when tackling users' interaction in a mobile multimodal environment is indeed huge and makes a rigorous quantitative analysis very difficult to pursue.

In this paper we tackle one of the challenges in designing multimodal interfaces for small mobile devices, that has not been quantitatively investigated. The issue is that while different tasks are traditionally considered suitable to a specific interface modality, mobility generally demands many rapid instantaneous actions that take place within different and changing circumstances that may make another modality more suitable. For example, dialing a number is expedient by using the keys or browsing the contact list, but when in a car – where eyes should be focused on the road, voice dialing might be a preferred method.

The issue of choosing a mode between several options in a dynamic usability environment on a small device was not sufficiently studied. We propose a model that can quantify user interaction and may enable a dynamic and adaptive choice of modality in the environment of a mobile multi-modal device.

## Methods

In our model for modality choice, we focus on a specific set of user actions which we denote by the term "Micro-actions". Many tasks of a mobile device involve relatively long physical action or a series of shorter actions. For example, updating a contact, creating a text message or navigating to an address, can take several minutes to accomplish and demands several consecutive actions from the user. The Micro-Actions, in contrast, are those linked with tasks like dialing, which should be fast: on the scale of seconds at most. Those tasks are contemplated by the user on a spur of a moment and require a result within a minimal latency. Those are the actions that most need an efficient and fast decision on which modality to use.

A micro-action can be described by 4 dimensions :

*action-trigger, interaction-environment, interaction-method* and *action complexity.*

An *action-trigger* is the type of the desired result. This axis is ever expanding as the functionality of mobile devices grows. Traditional action triggers in cell phones are call, texting a message, navigate to location, etc.

An *interaction-environment* is the scenario associated with the task: sitting, walking, driving and can be defined by user's location and motion, and to a certain amount, by the time in which the action was performed.

*An interaction-method* is one of a set of interaction types which are available for each action trigger. This axis currently consists of voice, keys, touch, gestures,

and again, is constantly expanding as technology progresses.

*The action complexity* is the complexity associated with the action till the desired result is achieved and takes into account both the duration of the micro action and the number of repetitions or corrections needed till an adequate result is achieved. This complexity can also be expressed as the time (real or perceived one), or the latency from the moment a user decides to perform an action until the achievement of the required task, or the "thought to action to result" process time.

The ultimate goal is a mapping of each micro-action to a target function of optimal modality. This process should minimize the *action complexity* of each Micro-action.

We proceed to define a measurement method that could compare *action complexities* for different actions, defined by their *action-triggers (T), interaction-environments (E) and interaction-methods (M)*.

An *action complexity* in constructed from the following building blocks:

1. $T1_{(T,E,M)}$ = Time to perform an action once. The action can be, for example, dialing a phone number (*trigger*=Dialing) in a car (*environment* =Car), using voice (*method*=Voice). Hence (T,E,M)=(D,C,M).

2. $P1_{(T,E,M)}$ = Probability of a correct result in a single trial. For the latter example: the probability that all digits dialed were correct.

3. $TFb_{(T,E,M)}$ = Duration of the system's feedback. In our example: a TTS of the number recognized or its presentation on a screen.

4. $TC_{(T,E,M)}$ = Time to correct an incorrect result (either to perform the action all over again or to correct a part of it).

5. $PC_{(T,E,M)}$ = Probability of getting the desired result after a correction.
(Note that in many cases performing the action and performing correction is the same action)

6. $TT_{(T,E,M)}$ = The total time to complete a micro action.

Under the assumption that repeating errors are statistically independent, TT (t,e,m) can be calculated as follows:

$$TT_{(T,E,M)} = T1_{(T,E,M)} + (1-P1_{(T,E,M)})( (TFb_{(T,E,M)} + TC_{(T,E,M)}) + (1-PC_{(T,E,M)}) (TFb_{(T,E,M)} + TC_{(T,E,M)}) + (1-PC_{(T,E,M)})^2 (TFb_{(T,E,M)} + TC_{(T,E,M)}) \ldots + (1-PC_{(T,E,M)})^n (TFb_{(T,E,M)} + TC_{(T,E,M)})) = T1_{(T,E,M)}+(1-P1_{(T,E,M)}) (TFb_{(T,E,M}+TC_{(T,E,M)}) / PC_{(T,E,M)}$$

**(1)**

This, however, is a simplistic "cold" calculation, and does not truly model the actual "complexity" associated with the action: Studies have shown that human's perception of time can be much longer than actual when the interaction is unpleasant or annoying [7, 8]. In this case, the annoying circumstance is the need to correct an error. The time perceived by the user who needs to correct over and over again can be longer than the actual time. We attempt to model this by replacing this term with a new one: a *perceived* time ($TPC_{(T,E,M)}$). This term should take into account the fact that a great amount of corrections are annoying even if in fact the time to their completion is relatively short and thus should be represented by a longer time than the actual one.

To quantify this phenomenon, we define a factor α: the "extra" perceived time which results from an iteration of, or the need to correct, an action, compared to the time it took originally to perform that action ($T1_{(T,E,M)}$).

The updated perceived correction time is the maximum of this new term and the previous sum of the feedback and correction time:

$$TPC_{(T,E,M)} = \max (TFB_{(T,E,M)} + TC_{(T,E,M)}, \alpha T1_{(T,E,M)})$$

**(2)**

The total perceived time for an action ($TPT_{(T,E,M)}$) is thus computed as follows:

$$TPT_{(T,E,M)} = T1_{(T,E,M)} + (1-P1_{(T,E,M)}) (TPC_{(T,E,M)}) / PC_{(T,E,M)}$$

**(3)**

We tested our formulae on different micro-action examples. The numbers associated with the different parameters were derived from users' experiments reported in the literature (i.e. [2]).

The annoyance factor α defined here was not measured in previous users experiments. We plan to measure it in our future experiments. In order to complete the calculations, however, we arbitrarily set α as 1.5. This assumption imply that each time a user

repeats an input to correct a mistake, it is perceived as 50% longer than the original, first input time.

**Results**

The first calculations were done for continuous digit dialing while driving in a car. This environment favor voice dialing since both hands should be on the wheel and both eyes on the road. The Micro-action parameters are thus a digit dialing trigger (D), a car environment (C) and a voice mode (V). The user utters a sequence of 10 digits, which takes on average 3 seconds ($T1_{(D,C,V)}$ = 3 Sec.). In a quiet car, the probability of getting the number right, $P1_{(D,C,V)}$ , can be 0.9. The feedback can be the system repeating the number recognized to the user is similar to the user's and hence is $TFB_{(D,C,V)}$ = 3 Sec. The correction is performed using similar method: repeating the whole sequence of digits again and the probability of correct recognition is still the same: 0.9. Therefore $TC_{(D,C,V)}$ = $T1_{(D,C,V)}$ and $PC_{(D,C,V)}$ = $P1_{(D,C,V)}$ . Using eq.1, the time for completing an action of voice digit dialing ($T_{(D,C,V)}$) is:

$TT_{(D,C,V)}$ =
$T1_{(D,C,V)}$ + (1-$P1_{(D,C,V)}$) ($TFB_{(D,C,V)}$ + $TC_{(D,C,V)}$) /
$PC_{(D,C,V)}$ = 3 + (1 – 0.9) (3 + 3) / 0.9 = 3.67 sec

The perceived time (eq. 2) in this case remain the same, since the operation needed for correction includes a sum of system feedback and user correction which is twice the original input time:

$TPC_{(D,C,V)}$ = max ($TFB_{(D,C,V)}$ + $TC_{(D,C,V)}$ , α $T1_{(D,C,V)}$)= max (6, 1.5 * 3) = 6

A change in the environment, i.e. a noisier car or an open window ("e"="NC"), will mainly affect the probability of correct recognition. If in this case the recognition drops to $P1_A$ = 0.6, eq. 1 yields:

$TT_{(C,NC,V)}$ = 3 + (1 – 0.6) (3 + 3)/0.6 = 7 sec

And again eq. 3 yields a similar result: $TPT_{(D,NC,V)}$ = $TT_{(D,NC,V)}$

The environment change thus doubles the Micro-action completion time and is, of course, problematic from the usability point of view.

If the environment further deteriorates and/or the recognition engine is of poorer quality the time to complete the action grows exponentially. For example, if the probability of correct recognition drops to $P1_A$ = 0.3, the time till completion will be:

$TT_{(D,NC,V)}$ = $TPT_{(D,NC,V)}$ = 20 sec, which, of course, is unacceptable in real life usage.

In the latter example, the feedback and correction times were longer than the "annoyance" term. An example for a different case can be a variant of the user interface mode, which provides a visual, head-up display feedback that is and verified by the user in 0.5 sec. The annoyance factor now shifts the perceived time: $TC_{(D,C,V)}$ = 3 + 0.5 sec, and hence:

$TPC_{(D,C,V)}$ = max ($TFB_{(D,C,V)}$ + $TC_{(D,C,V)}$, α $T1_{(D,C,V)}$) = max (3+0.5, 1.5 * 3) = 4.5 sec

$TT_{(D,C,V)}$ = 3 + (1 – 0.6) (4.5)/0.6 = 6 sec

(Compare to actual time which is 5 sec.)

Let us now consider a different system which allows only discrete digit dialing, in which each digit is followed by a feedback and an option of correction. The time for uttering one digit ("1D") is now 0.5 and is similar to the time needed for correction: $TC_{(1D,C,V)}$ = $T1_{(1D,C,V)}$ =0.5. Likewise, the probability for correct recognition in the first time and after a correction are similar and higher than a 10 digit's sequence recognition: $PC_{(1D,C,V)}$ = $P1_{(1D,C,V)}$ =0.95

The Action performance time per single digit will be

$TT_{(1D,C,V)}$ = 0.5 + 0.05 (0.5 + 0.5) / 0.95 = 0.55 sec

And if 10 such actions are performed for a 10-digits number the concatenated action time will be: $TT_{(D,C,V)}$ = 10 $TT_{(1D,C,V)}$ = 5.5 sec

A Comparison of this number to the former systems of continuous 10 digits dialing yields better performance if the continuous system's recognition rate is lower than 70%.

**Discussion**

The usability issue discussed in this paper focuses on tasks which require an immediate response which we denoted Micro-Actions. The challenge in these scenarios, which are common in mobile devices, is to accomplish a minimal time from "thought to execution". We proposed a model that could quantify the latency from "desired result" to "completed result" in a multi-modal user interface, for a given environment.

The main contribution in our analysis is the new look on the error correction process. We propose that the actual time needed for corrections, as measured in previous works, might not be accurate when a user's satisfaction is considered. Even very fast corrections, when they come in abundance, may seem very slow to a user due to the annoyance associated with it. This observation led to a definition of a *perceived* time which may be, in such cases, longer than the actual measured time. A consideration of *perceived* time is essential within our context of "Micro-

actions" where an instantaneous completion of a desired task is required by the user.

We demonstrated several calculations of perceived time for micro-actions using voice interface. This model for "thought till results" latency can be used to quantitatively analyze the value and usability of different modalities, environments and system recognition limitations.

An advanced multi modal system can learn its user's behavior in various conditions and add this information to the modality choice process: When the user decides to perform a certain action such as calling a number, the system can automatically suggest a modality given not only the task and environment, but also additional data like previously recorded user behavior.

This approach can thus help in the design of a successful and natural flow user interaction, needed in the evolving mobile environment and its increasing requirements.

**References**


1. Jaimes, A. and N. Sebe, *Multimodal human computer interaction: A survey.* Computer vision in human-computer interaction, 2005: p. 1-15.
2. Arif, A.S. and W. Stuerzlinger. *Predicting the cost of error correction in character-based text entry technologies*. 2010. ACM.
3. Suhm, B. *Empirical evaluation of interactive multimodal error correction*. 1997. IEEE.
4. Oulasvirta, A., et al. *Interaction in 4-second bursts: the fragmented nature of attentional resources in mobile HCI*. 2005. ACM.
5. Bernhaupt, R., et al., *Model-Based evaluation: A new way to support usability evaluation of multimodal interactive applications.* Maturing Usability: Quality in Software, Interaction and Quality, Springer, London, 2007.
6. Ling, R. *The length of text messages and use of predictive texting: Who uses it and how much do they have to say?* 2006.
7. Rajamony, R. and M. Elnozahy. *Measuring client-perceived response times on the WWW*. 2001.
8. McMillan, S.J. and J.S. Hwang, *Measures of perceived interactivity: An exploration of the role of direction of communication, user control, and time in shaping perceptions of interactivity.* Journal of Advertising, 2002: p. 29-42.