A Study in Efficiency and Modality Usage in Multimodal Form Filling Systems

Manolis Perakakis*, *Student Member, IEEE,* and Alexandros Potamianos, *Member, IEEE*

Abstract

The usage patterns of speech and visual input modes are investigated as a function of relative input mode efficiency for both desktop and personal digital assistant (PDA) working environments. For this purpose the form-filling part of a multimodal dialogue system is implemented and evaluated; three multimodal modes of interaction are implemented: “Click-to-Talk”, “Open-Mike” and “Modality-Selection”. “Modality-Selection” implements an adaptive interface where the system selects the most efficient input mode at each turn, effectively alternating between a “Click-to-Talk” and “Open-Mike” interaction style as proposed in [1]. The multimodal systems are evaluated and compared with the unimodal systems. Objective and subjective measures used include task completion, task duration, turn duration and overall user satisfaction. Turn duration is broken down into interaction time and inactivity time to better measure the efficiency of each input mode. Duration statistics and empirical probability density functions are computed as a function of interaction context and user. Results show that the multimodal systems outperform the unimodal systems in terms of objective and subjective criteria. Also users tend to use the most efficient input mode at each turn; however, biases towards the default input modality and a general bias towards the speech modality also exists. Results demonstrate that although users exploit some of the available synergies in multimodal dialogue interaction, further efficiency gains can be achieved by designing adaptive interfaces that fully exploit these synergies.

EDICS Category: SLP-SMMD

Index Terms

Graphical user interfaces, Input modality selection, Mobile multimodal interfaces, Speech communication

I. Introduction

The emergence of powerful mobile devices such as personal digital assistants (PDAs) and smart-phones, raises new design challenges and constraints that could be better addressed by a combination of more than one modalities. As defined in [2], [3], [4] multimodal interfaces process two or more combined user input modalities such as speech, pen, touch, manual gestures, gaze, head and body movements in a coordinated manner with multimedia system output. Efforts to build multimodal interfaces for PDAs are described in [5], [6], [7]. These systems inspired by Bolt’s “Put that there” [8] prototype mainly focus on map applications which can use speech and pen (gesture) input in a simultaneous fashion. Although map-based applications exemplify the advantages of multimodal vs. unimodal interaction by maximizing the synergies between modalities, information-seeking applications that involve form-filling are much more common on the PDA application environment, e.g., travel information and reservation, financial information and transactions, entertainment information. Typical form-filling applications used in MiPad [9], [10] a multimodal PDA prototype use “Tap and Talk” (a.k.a.“Click-to-Talk”) sequential multimodality (as opposed to concurrent multimodality [11]), i.e., only one input mode is active at each interaction turn. In this paper, we focus on information-seeking multimodal systems with speech and pen input, and investigate a variety of multimodal interaction modes in addition to “Click-to-Talk”.

It is widely supported that voice user interfaces (VUI) and graphical user interfaces (GUI) when combined to create a multimodal system offer high complementarity for most applications [12], [13], [14], [15]. As far as input

Manolis Perakakis is with the Dept. of Electronics & Computer Engineering, Technical Univ. of Crete, Chania 73100, Greece; email: perak@telecom.tuc.gr; tel:+30-28210-37258; fax:+30-28210-37542.

Alexandros Potamianos is with the Dept. of Electronics & Computer Engineering, Technical Univ. of Crete, Chania 73100, Greece; email: potam@telecom.tuc.gr; tel:+30-28210-37221; fax:+30-28210-37542.

This work was partially supported by the EU-IST-FP6 MUSCLE network of excellence and the GSRT 03ED375 PENED research project. The PENED research project is co-financed by E.U.-European Social Fund (75%) and the Greek Ministry of Development-GSRT (25%). The authors wish to thank Prof. Aikaterini Mania for valuable discussions and comment on evaluation issues, Michail Toutoudakis and Fotis Garitos for project involvement and all the users who participated in the evaluation.

is concerned, GUI interfaces have low error rates and offer easy error correction. Although speech is not error-free, it may be more efficient for relatively high speech recognition accuracy and high verbosity. It is also considered the most natural type of input compared to other modalities such as GUI. As far as output is concerned, visual output is fast (parallel) compared to much slower (sequential) speech output. Thus, multimodal systems that combine visual and speech interfaces can potentially become more efficient in terms of time to complete a task by taking advantage of: (i) "input modality choice" synergy, i.e., the user (or system in an adaptive user interface) chooses the most appropriate input modality for each context (ii) "visual-feedback", i.e., the more efficient cognitive processing of visual compared to auditory information, (iii) "error-correction" synergy, i.e., correcting errors of the VUI via the GUI [16], (iv) "concurrent multimodality" synergy, i.e., using both spoken and visual input in the same turn to more efficiently complete a task. Although (iii) and (iv) can be thought of as a special case of (i), we mention them explicitly because of their importance and recent research interest in investigating these type of synergies.

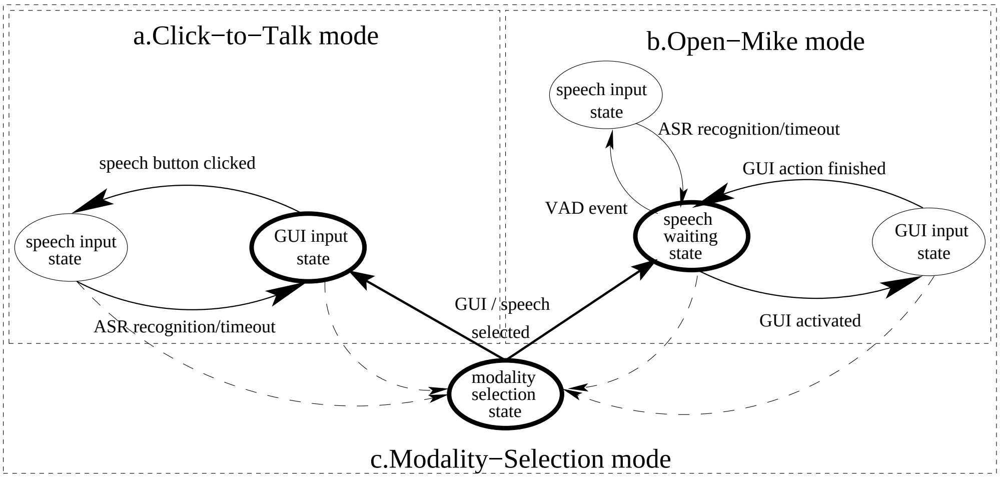


Fig. 1. State diagrams of the three multimodal modes : "Click-to-Talk", "Open-Mike" and "Modality-Selection" (from [17]).

A fundamental issue one has to consider when building multimodal interfaces is the suitability of various input methods for different tasks and subtasks [18]. For example, in [19] the authors compared data entry of isolated word Automatic Speech Recognition (ASR) with keyboard/mouse interfaces for three different data entry tasks: textual phrase entry, selection from a list and numerical data entry. Results indicated that speech input is faster for textual phrase entry if typing speed is below 45 words/minute. It is also faster for list selection when the list contains more than 15 items but offers no advantage over keypad or mouse for numerical data entry. Combining multiple modalities efficiently is a complex task and requires both good interface design and experimentation to determine the appropriate modality mix. Few guidelines exist for selecting the appropriate mix of modalities [20], [21], [22]. It is often the case when designing multimodal user interfaces that the developer is biased either toward the speech or the visual modality. This is especially true, if the developer is speech-enabling an existing graphical user interface (GUI)-based application or building a GUI for an existing speech-only service.

Another issue that is not thoroughly researched is the design of multimodal turn-taking and the selection of the most efficient interaction mode. Should users be allowed to interact as in traditional spoken dialogue systems (SDS) where a voice-activity detector allows the user to barge-in and speak at any moment (commonly referred as an "Open-Mike" interaction mode), should the user be constrained as in the GUI paradigm to press a button to activate the speech recognizer ("Click-to-Talk"), or should either interaction modes be used were appropriate.

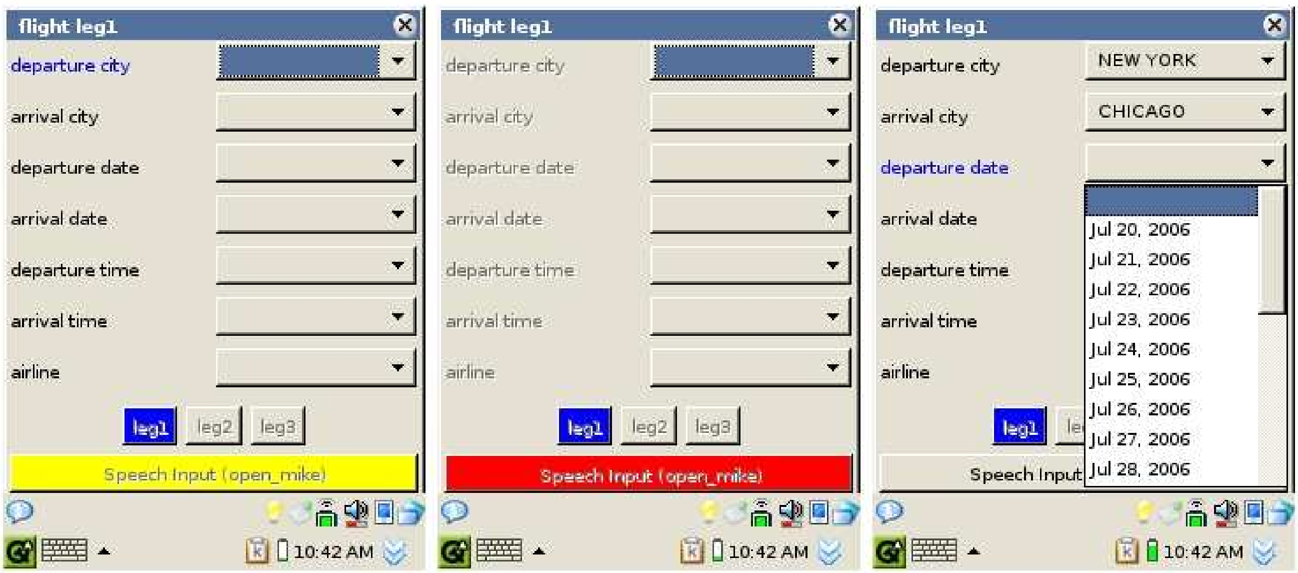
Our goal in this paper is to investigate input modality usage from the user point of view and to better understand efficiency considerations and user biases in input mode selection. Such information would be valuable for user modeling and multimodal dialogue system design in general. We have implemented and evaluated a travel reservation form-filling multimodal dialogue system for both desktop and PDA environments. The desktop system combines keyboard, mouse and speech input while the PDA system combines pen and speech input. Three multimodal modes were implemented, namely: "Click-to-Talk", "Open-Mike" and "Modality-Selection". For "Click-to-Talk" interaction, GUI is the default input modality while for "Open-Mike" interaction, speech is the default input modality. "Modality-Selection" is a mixture of the other two multimodal modes. The multimodal systems are evaluated and compared with the unimodal systems ("Speech-Only", "GUI-Only"). Our goal is not only to compare the efficiency and the objective metrics among the different systems, as is typically done in the literature, but to also measure the various factors that could affect the efficiency and modality choice by the user. For this purpose, we compute interaction and inactivity times within a turn to better understand the effect of input modality on interface efficiency. In addition, we measure modality usage for different levels of relative efficiency of the input modalities. General conclusions can be drawn from these experiments that can guide us through the multimodal interface design process.

Our work is based on the unimodal and multimodal systems described and informally evaluated in [17]. Minor enhancements to the multimodal interface have been carried out and an additional system (speech input/multimodal output) has been implemented and evaluated. The following are the main contributions of the paper compared to [17] and to the state-of-the-art:

1) A detailed evaluation of multimodal interaction modes and the comparison with unimodal modes both in the PDA and desktop environments.
2) The break down of the turn duration into interaction and inactivity time to better investigate modality synergies and usage patterns.
3) The investigation of the relationship between input mode efficiency and modality usage patterns to help design multimodal interfaces that are both efficient and adapt to user preferences.

The analysis and evaluation yields some obvious and some not-so-obvious results that can serve as guidelines in the design of multimodal form filling systems on the desktop and PDA.

The rest of this paper is organized as follows. In Section II, the unimodal and multimodal travel reservation systems are described. Evaluation methodology is presented in Section III. Objective and subjective evaluation results are presented in Section IV. A summary of the main results on input modality usage and their relevance to multimodal system efficiency and multimodal interface design are discussed in Section V. We conclude the paper with Section VI.



departure city NEW YORK
arrival city CHICAGO
departure date: Jul 20, 2006 / Jul 21, 2006 / Jul 22, 2006 / Jul 23, 2006 / Jul 24, 2006 / Jul 25, 2006 / Jul 26, 2006 / Jul 27, 2006 / Jul 28, 2006

Fig. 2. "Modality-Selection" interaction mode examples on the PDA. System is in "Open-Mike" mode in the first frame (speech button is yellow indicating waiting for input), receives user input "From New York to Chicago" during the second frame (speech button is red showing activity) and switches to "Click-To-Talk" mode in the third frame. The speech/pen input default mode is selected by the system in the first/third frame, respectively, due to the large/small number of options in the combo-box.

II. Unimodal and multimodal interaction

Our system is built using the multimodal spoken dialogue platform described in [23] for the travel reservation application domain (flight, hotel and car reservation). The multimodal system is implemented for both the desktop and PDA environments, with minor differences in the GUI design stemming from user interface consideration as explained next. For the purposes of this paper we have also used the unimodal spoken dialogue travel reservation

Bell Labs Communicator system described in [24], [23]. Our analysis focuses on the form-filling part, and ignores the result presentation and navigation part of the application.

The user can communicate with the system using pen and/or speech on the PDA, and using keyboard/mouse and/or speech on the desktop. Overall, five different interaction modes were implemented; two unimodal ones, namely, "GUI-Only" (GO) and "Speech-Only" (SO), and three fully multimodal ones, namely, "Click-to-Talk" (CT), "Open-Mike" (OM) and "Modality-Selection" (MS). In addition, a sixth interaction mode with unimodal speech input and multimodal visual and speech output labeled "Open-Mike Speech-Input" was implemented. The various systems and interaction modes are described next.

A. Unimodal GUI interaction

The application GUI (see Fig. 2) is generated automatically from the application ontology and the interface specification as described in [23]. It depicts the application state, using a series of forms; each form contains attribute-value pairs, each employing label and text-field/combo-box components, respectively. Two versions of the GUI are implemented: a desktop version which allows for keyboard and mouse input (GUI uses both text-field and combo-box components) and a PDA version which only allows for pen input (GUI uses only combo-box components).

Since desktop GUI systems allow for both mouse and keyboard (text) input our desktop GUI implementation exploits both input types to allow for fast GUI interaction. The choice of using text field or combo-box for a certain attribute field, is based on efficiency considerations; that is the number of values of that attribute in the ontology specification. If the number of values is small; i.e., less than 25 values, a combo-box is used, otherwise a text-field (see Table III). This combination has been found to be the most efficient for our application. For the PDA GUI on the other hand, *all* data entry fields are implemented as combo-box components due to the slow text input methods available on such devices. The number of options available to the user in some of these combo-box components is quite large, e.g., 250 choices for the "hotelname" attribute. Note that attribute values in any combo-box appear sorted in alphabetically order.

The following features are common for both the desktop and PDA GUI: (1) ambiguity is shown as a pull-down box with a list of choices and highlighted in red, (2) error messages are represented in the GUI as pop-up windows, (3) fields and buttons that become inaccessible in the course of the interaction are "grayed out", and (4) the context (or focus) of the interaction is highlighted.

B. Unimodal speech interaction

The original Communicator uses the BLSTIP [25] telephony platform. To further develop and explore multi-modality features on the Communicator, a highly flexible audio platform was designed and implemented, which can be run on both desktop computers and mobile devices (for various OS). It implements both *Voice Activity Detection* (VAD) and *barge-in*, i.e., users speaking over system prompts. The detailed description of the platform is beyond the scope of this paper. The audio platform interfaces with Bell Labs recognizer [25] and the FreeTTS [26] synthesizer through network sockets.

The "Speech-Only" interface is identical to the one described in [23], [24], [1]. In brief, the spoken dialogue manager promotes mixed-initiative system-user interaction. All types of user requests and user input are allowed at any point in the dialogue, i.e., the full application grammar is active throughout the interaction. The system prompts are focused and try to elicit specific information from the user, e.g., the value of an attribute. Explicit confirmation is used only to confirm the values of the attribute at the form level, e.g., for all flight leg user supplied information. Implicit confirmation is used in all other cases throughout the interaction.

C. Multimodal interaction

Three different multimodal (MM) interaction modes have been implemented for combining the visual and speech modalities. The output interface is common for all interaction modes to allow us to better investigate the effectiveness of the "optimum" input modality mix. The visual output is identical to the corresponding "GUI-Only" mode.

Audio output prompts were significantly shortened compared with the unimodal "Speech-Only" case. In general, speech output was mainly used as a way to grab the attention of the user, emphasizing information already appearing on the screen. The speech interface was identical for all multimodal modes.

TABLE I

SUPPORTED INPUT AND OUTPUT MODALITIES IN THE IMPLEMENTED SYSTEMS.

	input modalities		**output modalities**	
system	**GUI**	**speech**	**GUI**	**speech**
GO	√	x	√	x
SO	x	√	x	√
OMSI	x	√	√	√
CT/OM/MS	√	√	√	√

The state diagrams of the three multimodal modes are shown in Fig.1. For “Click-to-Talk” GUI is the default input modality; for speech input, users have to click the “Speech Input” button. For “Open-Mike” speech is the default input modality and allows the user to switch to visual input by clicking on the GUI. “Modality-Selection” mode is a mixture of “Click-to-Talk” and “Open-Mike” and attempts to better balance the visual or speech input modalities. It is a simple version of the adaptive modality tracking algorithm proposed in [1].

Note that in all three multimodal modes only one modality is active at a time, i.e., the system does not allow for concurrent multimodal input[1]. In our current multimodal implementation, visual input is not allowed (GUI is “grayed-out”) while speech input is active. Also, for all multimodal modes, users are free to override the system’s proposed input modality, that is, use a modality other than system’s default, e.g., GUI input for “Open-Mike” mode. The functionality of each multimodal mode is discussed in detail next.

Multimodal interaction modes: The main difference between the three multimodal interaction modes is the default input modality. For “Click-to-Talk” interaction pen is the default input; the user needs to click the “Speech Input” button to override the default input modality and use speech input. For “Open-Mike” interaction, speech is the default input modality; the system is always listening and a VAD event activates the recognizer. Again the user can override by pressing with the pen anywhere on the GUI; “Modality-Selection” is a mix of the “Click-to-Talk” and “Open-Mike” interaction; the system switches between the two multimodal modes depending on efficiency considerations (the size of the current combo-box). For combo-boxes with less than 25 values, the system goes into “Click-to-Talk” mode and visual input is the default mode, otherwise the system goes into “Open-Mike” mode where speech input is the default. Speech input is faster compared to pen input for “long” combo-boxes on the PDA and the threshold of 25 options was chosen based on the input mode efficiency of the stereotypical user. The state diagrams of the three multimodal interaction modes is shown in Fig. 1.

In Fig. 2, examples from the “Modality-Selection” mode running on the PDA, are shown. Initially the interaction focus is on “departure city”, the speech modality is selected (over 25 options available) and the system goes to “speech waiting” state. User input “from New York to Chicago” activates the speech recognizer (VAD event) and the GUI becomes disabled (“speech input” state). Once recognition of the spoken utterance finishes, the GUI is updated and the modality is selected for the next turn (“modality selection” state). For the next turn, visual input is selected (focus is on “departure date” for which a combo-box with less than 25 choices is available) and the system goes to the “GUI input” state.

D. Speech-only input and multimodal output

For the purposes of completeness and to better investigate the effect of “visual feedback” in spoken dialogue interaction a system with limited multimodal capabilities was also implemented, namely “Open-Mike Speech-Input” (OMSI). The user is allowed only speech input while the system output is multimodal including both speech and visual feedback. OMSI interaction is equivalent to “Open-Mike” interaction with visual (GUI) input disabled. Alternatively OMSI can be seen as a “Speech-Only” system with visual feedback and shortened prompts. Note that the OMSI prompts are identical to the MM system prompts.

In Table I a summary of the systems is shown in terms of input and output modalities. Note that the three multimodal modes support all available input and output modalities.

[1] For information-seeking/form-filling multimodal applications this is not a major limitation.

III. Evaluation methodology

A. Evaluation setting

Evaluation for both desktop and PDA environments includes the "GUI-only" and the three multimodal modes. In addition, two speech-input modes were evaluated, namely "Speech-only" and "Open-Mike Speech-Input" (OMSI). Thus a total of 10 systems were evaluated. Evaluation took place in an office environment, with all software (spoken dialogue system, speech platform, visual interface) running on the same host computer for the desktop and speech-only systems. For the PDA system, evaluation took place with all the back-end software (spoken dialogue system, speech platform) running on the same host desktop computer and the front-end (visual interface) running on a Zaurus Linux PDA device. Note that OMSI evaluation took place in the desktop environment.

TABLE II

Evaluation scenarios

Scenario ID	**flight**			**hotel**	**car**
	leg1	**leg2**	**leg3**	**reservation**	**rental**
1	√	-	-	-	-
2	√	√	-	-	-
3	√	√	-	√	-
4	√	√	-	-	√
5	√	√	√	-	-

All systems were evaluated using five scenarios of varying complexity: one/two/three-legged flight reservations and round trip flights with hotel/car reservation. Table II summarizes the required attributes in each of the five scenarios. For example, in the first scenario the user is required to book a one-way morning flight from Las-Vegas to Miami on July 10th on Northwest airlines.

In Table III, the usage of attributes in each scenario as well as cumulative usage across scenarios is shown; attributes are ordered based on the number of available values in the grammar. We refer to the three attributes listed, namely "hotelname", "city" and "airline", that have more that 25 possible values as "long" attributes while the rest are referred to as "short". Note that the cumulative attributes usage across all scenarios is about the same for "long" and "short" attributes (20 vs 22). [2] Eight non-native English-speaking university students evaluated all systems on all five scenarios. All users had prior limited experience by participating in a previous evaluation of an older version of the system.

TABLE III

Attribute size and attribute usage for the five travel reservation scenarios

attribute name	**attribute size**	**scenario usage**					
		1	2	3	4	5	total
hotelname	250	0	0	1	0	0	1
city	135	2	3	3	3	3	14
airline	93	1	1	1	1	1	5
date	22	1	2	2	2	3	10
car type	15	0	0	0	1	0	1
car rental	10	0	0	0	1	0	1
time	9	1	2	2	2	3	10

The evaluation procedure is described next. First, users are given a short introductory document which explains the system functionality with emphasis on the modes to be evaluated. In order to familiarize users with the system before actual evaluation takes place, users are asked to complete a demo scenario using all different systems, for a maximum of 30 minutes. Finally evaluation takes place, by asking users to complete all five scenarios using all ten systems (a total of 50 sessions per user and 40 sessions per mode). Systems are evaluated in random order and logs for each session are saved for later processing by our analysis software (objective evaluation). Upon completion of all runs, an exit interview is conducted (user feedback and overall subjective evaluation).

[2] This means than on average, if modality selection is solely based on efficiency considerations, we expect that usage of speech and GUI input in multimodal modes will be roughly the same (more on this in Section IV).

B. Objective evaluation

Interface evaluation of multimodal dialogue systems is a fairly complex task and different metrics may be used to evaluate various aspects of such systems [27], [28]. Since we are mostly interested in computing the relationship between modality usage and relative efficiency of input modalities, two depended variables become of high importance: modality selection (GUI or speech) and user turn duration (that is the time spent in each turn, for user input to the system[3]).

In this paper, we focus in the form filling part of the interaction and most specifically on how the user provides attribute-value pairs to the system[4]. Other parts of the interaction such as confirmation questions, verification requests, and navigation among forms were not included in our current analysis. The main reason for this is that for the vast majority of these actions, users used GUI input, as it was clearly the faster and easier way to respond, e.g., click "Yes" on a dialog window. By excluding the navigation, confirmation and verification actions we avoid biasing the evaluation results.

Dialogue based form filling systems are turn based. Turn duration (refer to Fig. 3) is the sum of user and system processing/response duration. Interaction efficiency focuses on the first component, which in turn consists of user inactivity and interaction times as defined in the next section.

Based on user turn times, statistics like "average turn duration" (mean of turn time), "overall user times" (sum of turn time) and "number of turns" can then be computed for a certain factor (independent variable) of interest, such as the mode, the context and the user. The rest of the section discusses the projection of evaluation data to various factors in order to compute statistics for the two depended variables of interest, namely user time and modality selection.

Next we give a short summary of objective metrics used in this study along with their indented use :

- Input modality usage: Related to unimodal efficiency.
- Input modality overrides: How well does the multimodal interaction mode predict user input modality preferences.
- Inactivity/interaction times: Separate input efficiency (related to interaction times) from output efficiency and cognitive load (related to inactivity times).
- Context statistics: Relate input modality efficiency with modality usage.
- User statistics: Identify individual user patterns.

Note that in the analysis that follows, we will mainly focus on the PDA environment and refer to desktop results, only when important differences are found.

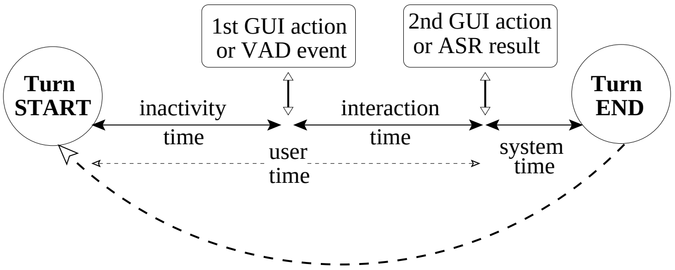


Fig. 3. Turn time decomposition to user and system time. Note that user time consist of inactivity and interaction time.

1) Modality selection and input modality overrides: The issue of modality usage is a focal point of our research. To answer this question a projection of our data on the modality factor is done. Specifically, we measure the usage of each input modality as a function of number of turns, and duration of turns attributed to each modality. Modality usage is also measured as a function of context, i.e., attribute for which input is expected, as discussed in Section III-B.3.

Another related measure is the number of input modality overrides, i.e., the number of turns where users preferred to use a modality other than the one proposed by the multimodal mode. Low number of overrides reveals that the

[3] This time also includes an overhead in the case of speech input (ASR overhead time), which has been found to be relatively small and is thus neglected by our current analysis.

[4] Note that error correction turns are included. We exclude from our analysis only turns that are responses to YES-NO questions such as "Is this a one way trip?" or "Is this correct?" (that occurs after filling out each form).

multimodal mode matches user's modality preferences and/or that the modality selection process is system-initiated for this user. A high number of overrides reveals a mismatch to user's modality preferences and/or a power-user that takes the modality selection initiative. The number of overrides is defined as the number of speech input turns for "Click-to-Talk" mode and as the number of GUI input turns for "Open-Mike" mode. For "Modality-Selection", the number of overrides is defined as the number of speech input turns for "short" attributes (where the system selects GUI input as default) plus the number of GUI input turns for the "long" attributes.

2) Turn duration, inactivity and interaction time: Duration statistics at the turn and task level are important factors, since in this work, efficiency is defined as being inversely proportional to task duration. In addition to measuring turn and dialogue duration in total and for each input modality, we also further refine turn duration into interaction and inactivity times (refer to Fig. 3). Inactivity time[5], refers to the idle time interval starting at the beginning of each turn, until the moment the user actually interacts with the system using GUI or speech input. During this interval, the user has to comprehend the system's response and state and then plan his own response after reading the scenario information. The response typically includes entering the system's requested information, using his preferred modality for that turn. We refer to this time as interaction time. By breaking the turn duration into interaction and inactivity time we can focus better on user input and system output, and thus separately investigate the input and output modality efficiency and usage patterns.

Inactivity time: For GUI input, the inactivity time is defined as the time interval between the turn start time and the moment the user clicks on the combo-box or starts writing in a text-field. For the case of speech input, inactivity time is defined as the time interval between the turn start time and the moment of a VAD event, that is the moment the audio subsystem has detected speech activity and starts sending speech samples to ASR. Note, that in the case of "Click-to-Talk" mode, we expect this time to be higher compared, e.g., to "Open-Mike", since the user has to also click the "Speech Input" button to start the audio recording first[6].

Interaction time: For GUI input, interaction time is defined as the time interval between the moment of the first GUI event and the moment the user selects the desired value in the combo-box or presses the TAB key after finishing text input entry (desktop case). For speech input, interaction time is defined as the time interval between the moment of the VAD event and the moment ASR result becomes available (this also includes an ASR overhead time as discussed earlier).

3) Context statistics: Context statistics refer to the objective measures regarding the attributes shown in Table III, also referred to as contexts. Given that the default modality in the "Modality-Selection" mode is chosen based on the number of available values for each attribute, modality usage and duration statistics as a function of context will help us better understand the relation between efficiency and modality choice.

In addition, to the traditional computation of the mean (and variance) of turn duration as a function of context, we also compute the empirical probability density functions (PDFs) of turn duration for each context. The empirical distributions are computed as a function of context and modality, for the interaction and inactivity time of each turn.

4) User statistics: User variability is another important issue that is investigated in this work. The relative efficiency of each modality is different for each user due to different speech recognition accuracy and prior experience with speech interfaces. Also users have different modality preferences (biases) that largely affect modality selection and performance. Individual user statistics also give us an idea of the degree of variability that users exhibit in making modality selection decisions and can help us to better understand the generality of the drawn conclusions on the relation between efficiency and modality usage.

C. Subjective statistics

The evaluation included an exit interview with a detailed questionnaire to measure the subjective opinion of each user for each system and modality. However, in the interest of space, and given that we are mostly interested in objective evaluation criteria (namely duration and modality usage) we only supply the overall subjective evaluation score given to the system by each evaluator. Correlations with objective metrics are also supplied. The subjective evaluation questionnaire used was similar to the one in [23].

[5] The term "inactivity" refers to the fact that the user *appears* inactive to the system.

[6] "Click-to-Talk" has voice activity detection enabled in this evaluation.

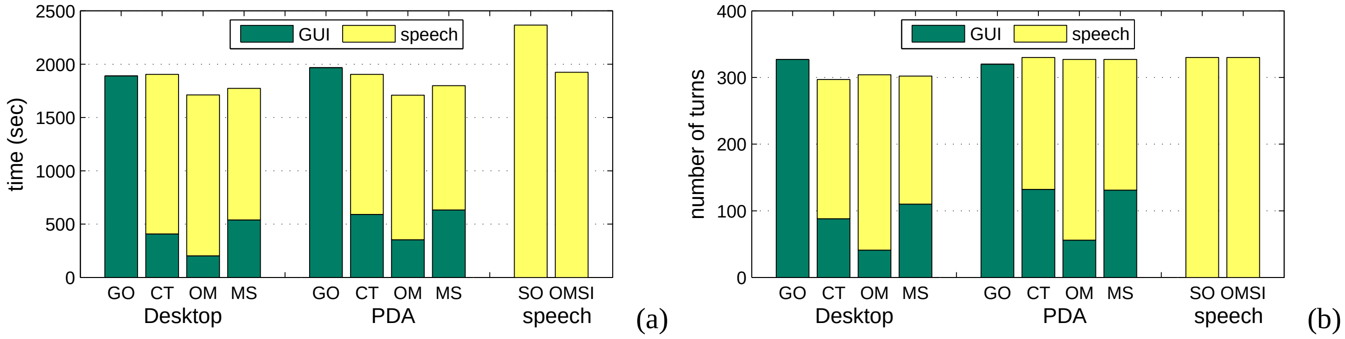


Fig. 4. Duration and turn cummulative statistics shown for each of the desktop PDA and speech-only systems summed over all scenarios: (a) total time to completion in seconds, (b) total number of turns. The color-codes for each system bar show the total time and number of turns for GUI and speech input respectively.

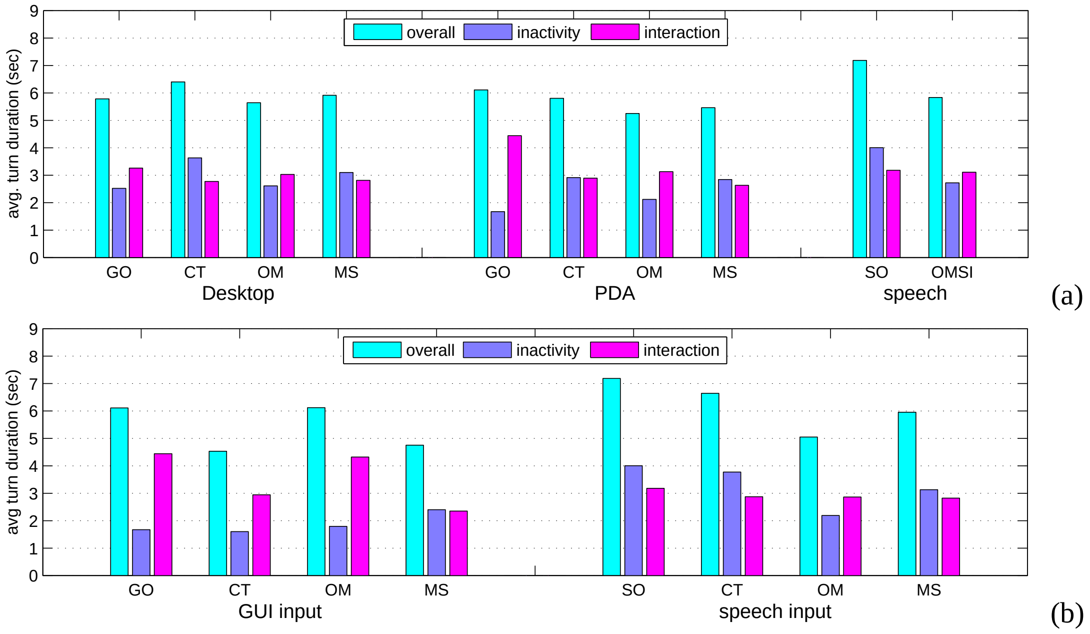


Fig. 5. (a) Average turn duration (in sec) for all ten systems (four Desktop, four PDA and the two speech-only systems) broken into inactivity and interaction times. (b) PDA inactivity and interaction times grouped by input type (GUI and speech) respectively. Note the “Speech-Only” (SO) system is also included as a reference.

IV. Evaluation Results

A. Mode performance

“User time” and “number of turns” for all ten systems (four for desktop, four for PDA and two speech input modes) are shown in Fig. 4(a) and Fig. 4(b) respectively.[7] Overall, “Speech-Only” is the less efficient mode. “Open-Mike Speech-Input” (equivalent to “Speech-Only” mode with visual feedback and reduced prompts) is much faster compared to “Speech-Only” mode; in fact, its efficiency is much closer to the efficiency of the multimodal modes rather than the efficiency of the “Speech-Only” mode. For both desktop and PDA environments, “Open-Mike” is the fastest mode closely followed by “Modality-Selection” mode and then by the slower “GUI-Only” and “Click-to-Talk” modes. Note that minor differences exist between desktop and PDA in “GUI-Only” mode efficiency, number of turns and GUI input usage (slightly higher for the PDA case - see Fig. 4(b)).

[7] Note that these results are normalized for 38 instead of 40 runs per system, to compensate the failure of 2 runs for “OMSI” and 1 run for “SO” systems. We did not include these outliers in our data to avoid biasing the results.

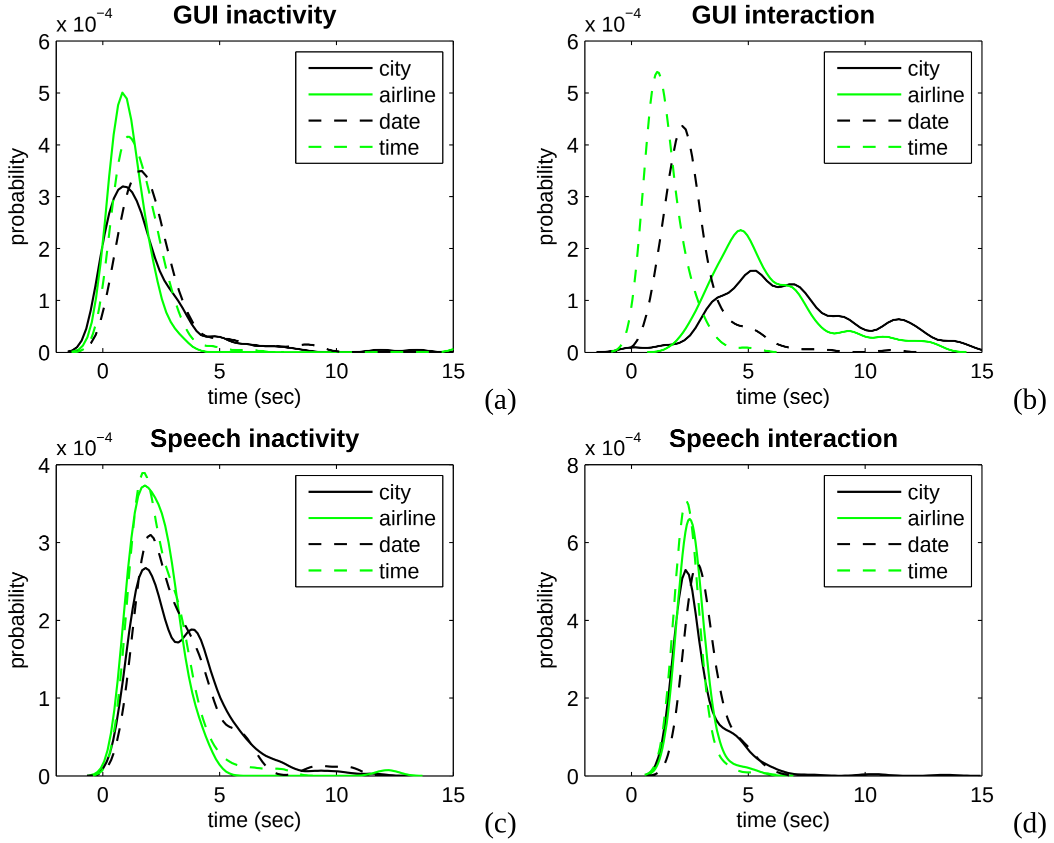


Fig. 6. Distributions of average turn duration in seconds broken down into inactivity/interaction times and input type (pen/speech) for the four most frequently-used contexts (city, airline, date, time). Results are cummulative for the four PDA systems (GO, CT, OM, MS). Distributions approximated using kernel density functions. (a) Avg. inactivity time distribution for pen input. (b) Avg. interaction time distribution for pen input. (c) Avg. inactivity time distribution for speech input. (d) Avg. interaction time distribution for speech input.

TABLE IV

SUMMARY OF MAIN OBJECTIVE STATISTICS.

	Overall		**Task Average**		**Turn Average duration(sec)**		
System	**completion rate(%)**	**GUI/Speech turns(%)**	**# turns**	**duration(sec)**	**inactivity**	**interaction**	**overall**
			Speech system evaluation				
SO	97.5	0/100	8.69	62.43	4.00	3.18	7.18
OMSI	95	0/100	8.50	59.09	2.72	3.10	5.83
			Desktop evaluation				
GO	100	100/0	8.68	50.10	2.52	3.26	5.78
CT	100	31/69	8.08	51.65	3.63	2.77	6.40
OM	100	32/68	8.10	45.70	2.61	3.03	5.64
MS	100	36/64	8.35	49.38	3.10	2.81	5.91
			PDA evaluation				
GO	100	100/0	8.50	51.95	1.67	4.44	6.11
CT	100	39/61	8.75	50.78	2.91	2.89	5.80
OM	100	18/82	8.93	46.82	2.12	3.13	5.25
MS	100	41/59	8.65	47.26	2.84	2.63	5.46

B. Turn duration, inactivity and interaction times

Fig. 5(a) shows average turn durations broken into interaction and inactivity times for all ten systems. We conducted ANOVA analysis for the four PDA and the two speech only systems (desktop systems are also shown in Fig. 5(a)). A within subjects ANOVA shows that the effect of system on inactivity ($F_{5,2014}$ = 83.78, $p < 0.001$), interaction ($F_{5,2014}$ = 33.98, $p < 0.001$) and overall times ($F_{5,2014}$ = 23.97, $p < 0.001$) are all highly significant. A post-hoc Tukey HSD test ($p < 0.05$) was performed to find any significant differences. For inactivity times, GO is the faster, followed by OM, then the CT, MS and OMSI systems (whose in-between differences are not statistically significant) and finally, SO which has the highest inactivity times. For interaction times, GO is by far the slower mode; there are no significant differences among the other systems, except for the MS that has the lowest interaction times. Finally, SO has the highest overall times, followed by GO, OMSI and CT, and finally the MS and OM systems. Note that the multimodal modes have shorter interaction times compared to "GUI-Only" and shorter inactivity times compared to "Speech-Only".

We next turn our attention to Fig. 5(b) that shows average turn durations broken into interaction and inactivity times and grouped by input type (GUI/speech) for the PDA system. SO (also shown in Fig. 5(b)) and OMSI systems are also included in the ANOVA analysis. A within subjects ANOVA shows that the effect of system/input on inactivity ($F_{8,1990}$ = 74.25, $p < 0.001$), interaction ($F_{8,1990}$ = 32.48, $p < 0.001$) and overall times ($F_{8,1990}$ = 23.32, $p < 0.001$) are all highly significant. A post-hoc Tukey HSD test ($p < 0.05$) was performed to find any significant differences.

For pen (GUI) input (left part of Fig. 5(b)), MS inactivity times are higher compared to GO, CT and OM whose in-between differences are not significant. For speech input (right part of Fig. 5(b)), SO and CT have the higher inactivity times, followed by MS, then OMSI and finally OM. All differences are significant except for SO and CT. For pen input, all interaction times differ, except for GO and OM (whose estimate is based only on 66 inputs); for speech input however there are no significant differences among the systems.

Furthermore, note the short inactivity times and varying interaction times for GUI input shown at Fig. 5(b). Inactivity times are short (compared to speech input inactivity times) and also roughly the same for all modes, except for the "Modality-Selection" mode. Interaction times vary considerably; they are very high for "GUI-Only" (no input modality choice) compared to multimodal modes. Note that for GUI input "Modality-Selection" has the highest inactivity but lowest interaction times.

As far as speech input is concerned, one can note almost identical interaction times for the three multimodal modes but highly varying inactivity times. "Open-Mike" has shorter speech inactivity times compared to "Click-to-Talk", while "Modality-Selection" inactivity times are approximately the average of the other two modes. Comparing GUI and speech input, it is evident that inactivity times are much shorter for GUI input compared to speech input (GUI click vs VAD event).

Table IV shows a summary of objective statistics for the current evaluation. The statistics are reported for all ten systems evaluated (two for speech only systems, four for desktop and PDA environments). Metrics include task completion rate, percent of GUI/speech turns, task related statistics such as average number of turns per task and average task duration. Turn related statistics such as average turn duration are also reported along with the break down into inactivity and interaction parts.

TABLE V

SPEECH INPUT CONTEXT STATISTICS: CONCEPTS PER TURN (VERBOSITY) FOR THE THREE PDA MULTIMODAL SYSTEMS AND AVERAGED % CONCEPT ACCURACY FOR FOUR CONTEXTS.

	CT	**OM**	**MS**	
context	**verbosity**			**% concept accuracy**
city	1.71	1.52	1.54	92
date	1.35	1.31	1.34	65
time	1.05	1.00	1.00	92
airline	1.00	1.00	1.00	88

C. Context statistics

Fig. 6 shows PDA inactivity and interaction time distributions grouped by input type (GUI or speech) for the four most frequently used attributes. Note that inactivity times for speech input are higher compared to GUI ones (Fig. 6(a) and Fig. 6(c)). As shown in Fig. 6(b) GUI interaction times clearly depend on attribute (combo-box) size, as expected, while speech interaction times (Fig. 6(d)) are similar for all attributes. Finally, interaction times for

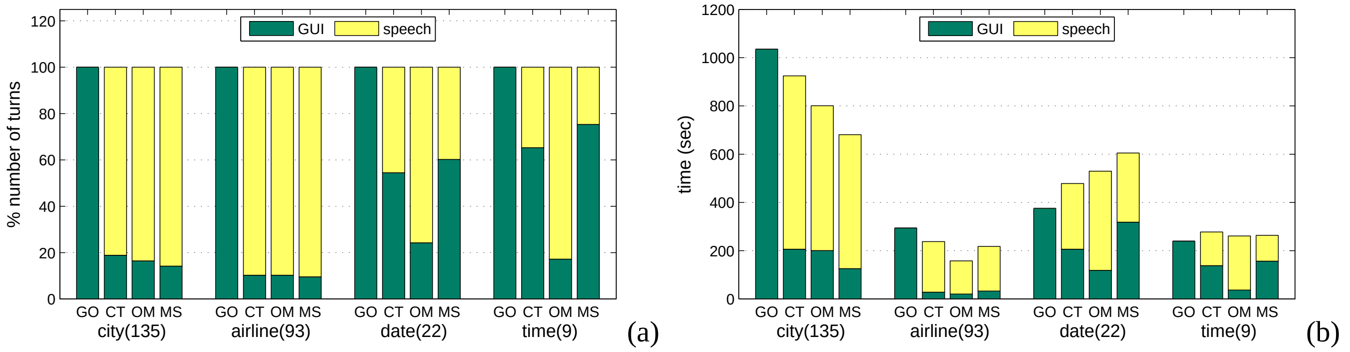


Fig. 7. PDA context statistics for the four most important attributes (a) percent number turns and (b) overall user time.

speech input are considerably shorter compared to GUI ones for the case of city and airline attributes but slightly longer for date and time attributes.

Table V shows the average number of concepts provided per turn (speech verbosity) for the various contexts (expected attribute input) grouped by system. In addition, the concept accuracy is shown defined as the percent of concepts recognized correctly by the system over the total number of concepts uttered by the user. Note that for the city and date attributes verbosity is high, e.g., for the city case users usually provide both departure and arrival city e.g., "From Las Vegas to Miami". The same holds for date, e.g., "July 25th in the morning". Also note that concept accuracy is high (about 90%) for all attributes except for date[8].

Fig. 7(a) shows input modality selection (% number of turns) for the four most frequently used attributes, sorted by size (e.g., 135 for city attribute). Speech usage is fairly high for "long" attributes (between 80% and 90%) and mode-independent. For "short" attributes on the other hand, speech usage is clearly mode-dependent i.e., for the time attribute it is 80% for "Open-Mike", 35% and 25% for "Click-to-Talk" and "Modality-Selection" respectively.

As shown Fig. 7(b) (user times for the same four attributes) multimodal interaction for all three modes is much faster compared to "GUI-Only" mode for "long" contexts. For "short" contexts (date and time), however, multimodal modes perform worse compared to "GUI-Only" mode.

D. Input modality overrides

Fig. 8 shows the (%) number of default input modality overrides for the three PDA multimodal modes; the four most important attributes are shown, sorted by size. For "Click-to-Talk" (CT) the number of overrides (use of speech instead of GUI input) is very high for "long" attributes where users preferred to override default GUI modality in favor of speech. For "Open-Mike" (OM) the number of overrides is the lowest. Very few overrides occur for "long" attributes and slightly more for "short" ones. Finally, although "Modality-Selection" (MS) has fairly low percent of overrides (use of GUI instead of speech input) for "long" attributes, percent of overrides for "short" attributes (use of speech instead of GUI input) is higher (between those of CT and OM). As a result "Modality-Selection" has slightly more overrides compared to "Open-Mike". Overall, OM has the least number of overrides, closely followed by MS and then CT where a very high number of overrides occurs.[9]

E. User statistics

Fig. 9(a) shows total duration for the three multimodal and the GO systems per user (PDA evaluation scenarios). Task duration per user is further broken down into duration of GUI-input and speech-input turns. Performance of "GUI-Only" mode highly varies between users. Also, for some users, multimodal modes are more efficient compared to "GUI-Only" (e.g., usr2, usr6, usr7) while for other users they are worse (e.g., usr8).

[8] The "date field" consists of two distinct words namely: month and day, e.g., "July 6". A recognition error for either word would result in an error for the "date" concept. Note that although the number of "legal" date values shown at the GUI are only 22, the speech recognition grammar is unconstrained allowing effectively 365 values.

[9] Note that override results were presented as a % of the input turns; one has to also consider the relative usage of the four attributes in Table III. Also, the cost of overrides may not be the same for all cases.

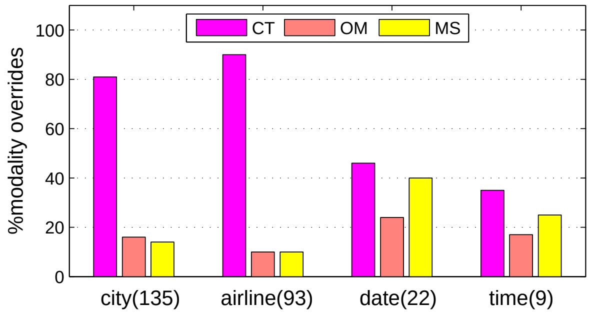


Fig. 8. Input modality overrides (%) for the three PDA multimodal modes (CT, OM and MS) grouped by attribute type (attribute size included in parentheses).

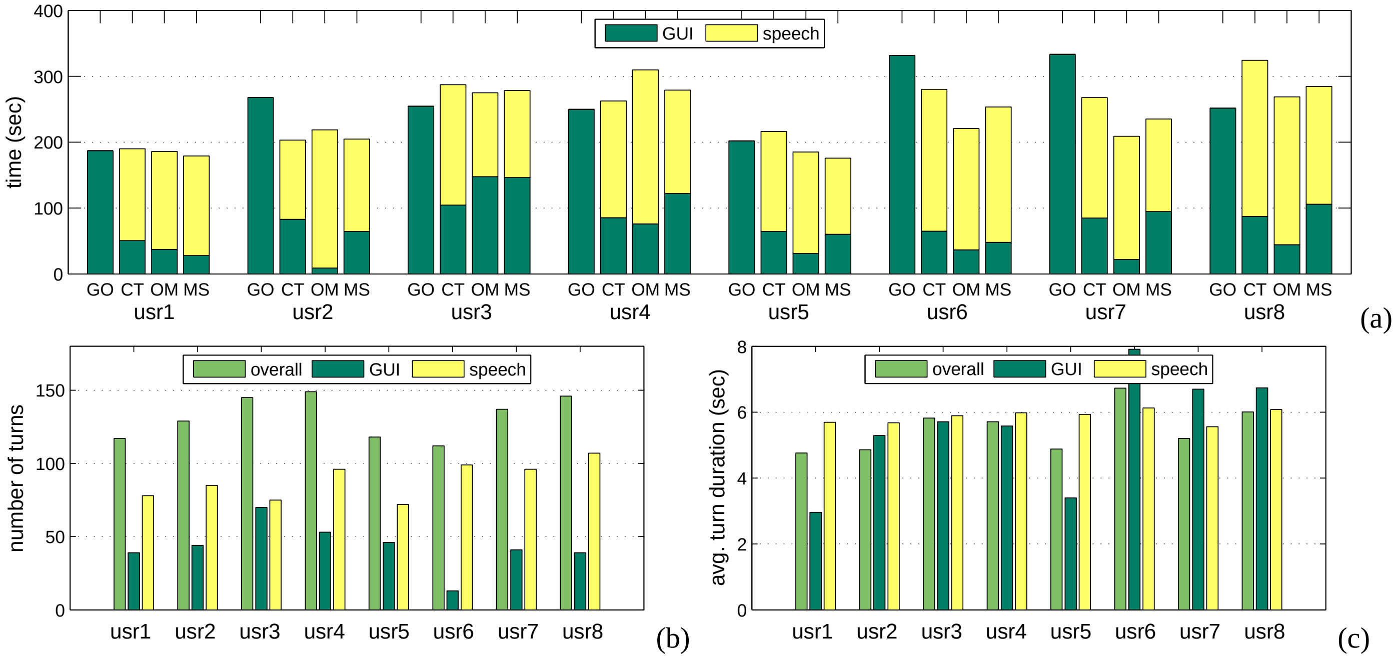


Fig. 9. PDA user statistics. (a) total time to completion for the multimodal and GO systems (b) sum of number of turns for the three multimodal modes (c) average turn duration for all three multimodal modes.

Fig. 9(b) shows number of turns and Fig. 9(c) shows average turn duration by averaging all three multimodal modes for the PDA case. For speech input turns, average turn duration is between 5 and 6 secs, while for GUI input turns, average turn duration is between 3 and 8 secs. Thus variability in duration (or variability in efficiency) among users is much higher for GUI input compared to speech input.

There is also high variability in the number of turns; the total number of turns depends on both the percentage of GUI and speech turns and the combination of speech verbosity and concept accuracy (thus the number of correct concepts per turn). We have found that concept accuracy varies between users from 75% to 94% while verbosity (number of concepts supplied per user turn) varies from 1.05 to 1.52. Note that some users (usr1, usr5 and usr6) completed all scenarios with less than 126 turns (more than one *correct* concept per turn for speech input), while others needed considerably more turns.

F. Subjective statistics

In Table VI, the overall subjective evaluation scores are shown for all ten systems. In the last two rows the mean and standard deviation are also reported. The overall scores were supplied by the users after the exit interview. A within subjects ANOVA shows that the evaluated systems differ ($F_{9,63} = 6.08$, $p < 0.001$). A post-hoc Tukey HSD test ($p < 0.05$) was performed to find any significant differences.

TABLE VI

SUBJECTIVE EVALUATION RESULTS

Platform	**desktop**				**PDA**				**speech**	
System	**GO**	**CT**	**OM**	**MS**	**GO**	**CT**	**OM**	**MS**	**SO**	**OMSI**
Usr1	9	10	10	9	9	9	10	9	7	10
Usr2	8	7	10	10	10	10	10	10	6	10
Usr3	10	7	8	7	7	9	10	8	4	5
Usr4	6	7	7	8	7	8	9	8	6	5
Usr5	8	8	10	9	8	9	10	10	7	8
Usr6	6	10	10	10	9	10	10	9	7	8
Usr7	8	9	9	10	8	9	9	10	8	9
Usr8	8	9	9	10	7	8	8	9	8	9
Mean	7.88	8.38	9.13	9.13	8.13	9	9.5	9.13	6.63	8
StDev	1.28	1.13	1.07	1.12	1.07	0.76	0.73	0.83	1.29	1.83

Results show that while SO significantly differs with all other systems, OMSI only differs with the OM and MS for both desktop and PDA systems, which got the highest ranking overall. The only significant differences for the desktop environment are among GO with OM and MS systems and for the PDA environment among the GO and OM system. Further analysis shows that the correlation between time-to-completion and overall subjective evaluation scores for the ten systems is relatively high (-0.43).

V. SUMMARY OF MAIN RESULTS

The results in Fig. 5(a) clearly show the importance of having "visual feedback" in a spoken dialogue system. By incorporating visual output to OMSI the efficiency increases dramatically (inactivity time decreases by 1.3 secs) compared to the "Speech-Only" system. "Input modality choice" also plays an important role; note the decrease in interaction time between "GUI-only" and multimodal modes, e.g., MS for the PDA case. By offering the users the freedom to select the most efficient input modality in any given context, interaction time can be shortened considerably. This is especially true for the "long" attributes (city and airline) in the PDA case for which speech input is much faster compared to GUI input.

A. Multimodal interaction modes

Among the three multimodal systems the "Click-to-Talk" system is clearly the least efficient. This is due to inefficiencies of this mode for speech input (observe the high inactivity times in Fig. 5(b)) combined with the relatively high percent of speech usage (see Fig. 8).

From Fig. 5(b), we can observe that GUI input has on average lower inactivity times, while speech input has lower interaction times. Although speech is the most efficient in terms of input (interaction times), recognition errors and context switching incurs higher cognitive load to the user resulting in higher inactivity times for speech input.

The "adaptive" "Modality-Selection" system, which at each turn suggests to the user the most efficient input mode, has the shorter interaction times, however it typically has high inactivity times. This is due to the increased cognitive load that adaptivity incurs on the user; automatically switching between default input modes is sometimes inconsistent and confusing. This is a common problem of adaptive interfaces.

Given that speech input usage was much higher in our current evaluation compared to GUI input, it is no surprise that "Open-Mike" is faster than "Click-to-Talk"; "Modality-Selection" being a mixture of the other two multimodal modes, has efficiency that lies somewhere between the efficiency of the other two modes.

B. Modality usage patterns

The mode statistics results in Fig. 4(b) clearly show that the multimodal system biases the input modality usage (CT vs. OM). Users tend to use GUI input more often when it is the default input mode (in CT), compared to the OM system where speech in the default input mode.

In Fig. 6(b), we see that the mean interaction times for GUI input are shorter for attributes with fewer options in the combo box, as expected. For speech input, the PDFs shown in Fig. 6(d) are very similar for all attributes. Comparing the interaction times per attribute, it is clear that GUI input is more efficient for "time" and "date", while speech input is more efficient for "city" and "airline".

Based on the observation above and given the almost 50-50% balancing between "GUI-efficient" and "speech-efficient" attributes in the scenarios, one would expect a 50-50% input modality usage split between GUI and speech. However, the results show that for all multimodal systems speech input is used for over 60% of the turns. A possible explanation for this, is that we have an asymmetrically balanced situation; that is, although our scenarios are almost balanced in number of turns (number of "long" vs "short" attributes), the difference in unimodal efficiency (GUI vs speech) for the "long" and "short" attributes is not symmetrical. Difference in efficiency between GUI and speech is much higher for "long" attributes (in favor of speech) compared to "short" ones (in favor of GUI) as shown in Fig. 6(b) and Fig. 6(d). Additionally users are aware of the relation of GUI efficiency with attribute (combo-box) size; however such a relation is not clear for speech input.

Speech errors also affect input modality selection. This can be clearly seen in Fig. 7(a) where users use GUI input for "long" attributes, only to correct speech recognition errors.

Subjective results show that users prefer multimodal modes compared to unimodal ones. Users seem to value both the visual feedback (OMSI vs SO) and the input modality choice offered by the multimodal modes (multimodal vs unimodal). Although the correlation between user times and subjective scores is high, other factors also affect users mode preferences.

C. User variability

From Fig. 9 and Table VI it is clear that the user patterns vary significantly as far as unimodal efficiency, modality selection and subjective scores are concerned. The users display significant differences in efficiency for GUI input and (expected) differences in efficient for speech input (due to difference speech recognition error rates). The users also differ significantly in input modality usage and preferred interaction mode. The high variability in user patterns shows that a "stereotypical" modality selection model (such as the one implied by the MS interaction mode) might not model adequately user modality preferences, and that a model/system that adapts individually to each user might be necessary.

Overall, combining multiple modalities efficiently is a complex task that requires both good interface design and experimentation to determine the appropriate modality mix. From the analysis of the relative efficiency of the input modes and from the modality usage results it is clear that a relationship between input mode selection and mode efficiency exists but is not perfectly linear. More research is needed to quantify the nature of this relationship and to design adaptive modality selection algorithms that maximize efficiency without conflicting with user preferences.

VI. Conclusions

In this paper, we have evaluated two unimodal and three multimodal form-filling systems on the desktop and PDA environments. The objective evaluation metrics used included mode and task duration statistics. These objective metrics were also calculated on a per attribute basis and broken down into the interaction and inactivity part of a dialogue turn. This detailed evaluation yielded some obvious and not-so-obvious results that can help us better understand human-machine interaction for multimodal dialogue systems. Here are some important conclusions from our analysis: (1) Synergies between the speech and visual interaction modes exist in multimodal interfaces; among these synergies visual feedback (GUI output) and input modality choice play an important role. (2) When changing the relative efficiency of the input modes in multimodal interfaces, user input modality usage also changes; users tend to use the most efficient modality but biases also exist. (3) Multimodal and adaptive interfaces are almost always better in terms of shorter interaction times, but inactivity time may increase due to increased cognitive load of the user. Keeping these points in mind can help us design better multimodal systems.

Future work will focus on evaluating the unimodal and multimodal systems for varying levels of task complexity and unimodal interface efficiency (e.g., different speech recognition error levels). Through these experiments multiple measurement points for modality usage, unimodal and multimodal interface efficiency will be obtained; these results will help us better understand the relationship between efficiency, user satisfaction and input modality usage. We will also perform longitudinal studies [29] to investigate possible "novelty effects" and adaptation of user interaction patterns. By incorporating this knowledge into the multimodal dialogue system design process we aim at building adaptive multimodal interfaces that are natural, efficient and improve on the state-of-the-art.

REFERENCES

[1] A. Potamianos, E. Ammicht, and E. Fosler-Lussier, "Modality tracking in the multimodal Bell Labs Communicator," in *Proc. of Automatic Speech Recognition and Understanding Workshop*, St. Thomas, U.S. Virgin Islands, 2003, pp. 192–197.

[2] S. Oviatt, "Multimodal interfaces," in *The Human-Computer Interaction Handbook: Fundamentals, evolving technologies and emerging applications*, J. Jacko & A. Sears, Ed., pp. 286–304. Lawrence Erlbaum: New Jersey, 2003.

[3] D. B. Koons, C. J. Sparrell, and K. R. Thórisson, "Integrating simultaneous input from speech, gaze, and hand gestures," in *Proc. of Artificial Intelligence Workshop on Intelligent Multimedia Interfaces*, Menlo Park, CA, USA, 1993, pp. 257–276.

[4] A. Waibel, M. Tue Vo, P. Duchnowski, and S. Manke, "Multimodal interfaces," *Artif. Intell. Rev*, vol. 10, no. 3-4, pp. 299–319, 1996.

[5] P. R. Cohen, M. Johnston, D. McGee, S. Oviatt, J. Pittman, I. Smith, L. Chen, and J. Clow, "Quickset: multimodal interaction for distributed applications," in *Proc. of ACM International Conference on Multimedia*, Seattle, Washington, United States, 1997, pp. 31–40.

[6] M. Johnston, S. Bangalore, G. Vasireddy, A. Stent, P. Ehlen, M. Walker, S. Whittaker, and P. Maloor, "MATCH: an architecture for multimodal dialogue systems," in *Proc. of the 40th Annual Meeting on Association for Computational Linguistics*, Philadelphia, Pennsylvania, 2002, pp. 376–383.

[7] S. Dusan, G. J. Gadbois, and J. Flanagan, "Multimodal Interaction on PDA's Integrating Speech and Pen Inputs," in *Proc. of Eighth European Conference on Speech Communication and Technology*, Geneva, Switzerland, 2003, pp. 2225–2228.

[8] R. A. Bolt, "Put-That-There : Voice and gesture at the graphics interface," in *Proc. of Computer Graphics and Interactive Techniques*, Seattle, Washington, United States, 1980, pp. 262–270.

[9] X. Huang, A. Acero, C. Chelba, L. Deng, D. Duchene, J. Goodman, H. Hon, D. Jacoby, L. Jiang, and R. Loynd, "MiPaD: A Next Generation PDA Prototype," in *Proc. of the International Conference on Spoken Language Processing*, Beijing, China, 2000, pp. 33–36.

[10] L. Deng, K. Wang, A. Acero, H. W. Hon, J. Droppo, C. Boulis, Y. Y. Wang, D. Jacoby, M. Mahajan, and C. Chelba, "Distributed speech processing in MiPaD's multimodal user interface," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 8, pp. 605–619, 2002.

[11] L. Nigay and J. Coutaz, "A design space for multimodal systems: concurrent processing and data fusion," in *Proc. of the INTERACT '93 and CHI '93 conference on Human factors in computing systems*, Amsterdam, The Netherlands, 1993, pp. 172–178.

[12] J. Lai and N. Yankelovich, "Conversational speech interfaces," in *The Human-Computer Interaction Handbook: Fundamentals, evolving technologies and emerging applications*, pp. 698–713. Lawrence Erlbaum Associates, Inc., Mahwah, NJ, USA, 2003.

[13] M. Grasso, D. Ebert, and T. Finin, "The Integrality of Speech in Multimodal Interfaces," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 5, no. 4, pp. 303–325, 1998.

[14] P. Cohen and S. Oviatt, "The Role of Voice in Human-Machine Communication," in *Voice Communication Between Humans and Machines*, pp. 34–75. National Academy Press, Washington D.C., 1994.

[15] P. Cohen, M. Johnston, D. McGee, S. Oviatt, J. Clow, and I. Smith, "The Efficiency of Multimodal Interaction: a Case Study," in *Proc. of Fifth International Conference on Spoken Language Processing*, Sydney, Australia, 1998, pp. 249–252.

[16] B. Suhm, B. Myers, and A. Waibel, "Multimodal error correction for speech user interfaces," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 8, no. 1, pp. 60–98, 2001.

[17] "Blending speech and visual input in multimodal dialogue systems," in *Proc. of the IEEE/ACM Workshop on Spoken Language Technology*, Aruba, 2006, pp. 142–145.

[18] N. O. Bernsen and L. Dybkjaer, "Is Speech The Right Thing For Your Application?," in *Proc. of Fifth International Conference on Spoken Language Processing*, Sydney, Australia, 1998, pp. 3209–3212.

[19] H. Mitchard and J. Winkles, "Experimental Comparisons of Data Entry by Automated Speech Recognition, Keyboard, and Mouse.," *Human Factors*, vol. 44, no. 2, pp. 198–210, 2002.

[20] V. Bilici, E. Krahmer, S. te Riele, and R. Veldhuis, "Preferred Modalities in Dialogue Systems," in *Proc. of Sixth International Conference on Spoken Language Processing*, Beijng, China, 2000, pp. 727–730.

[21] L. M. Reeves, J. C. Martin, M. McTear, T. V. Raman, K. M. Stanney, H. Su, Q. Y. Wang, J. Lai, J. A. Larson, and S. Oviatt, "Guidelines for multimodal user interface design," *Communications of the ACM*, vol. 47, no. 1, pp. 57–59, 2004.

[22] K. Stanney, S. Samman, L. Reeves, K. Hale, W. Buff, C. Bowers, B. Goldiez, D. Nicholson, and S. Lackey, "A Paradigm Shift in Interactive Computing: Deriving Multimodal Design Principles from Behavioral and Neurological Foundations," *International Journal of Human-Computer Interaction*, vol. 17, no. 2, pp. 229–257, 2004.

[23] A. Potamianos, E. Fosler-Lussier, E. Ammicht, and M. Perakakis, "Information seeking spoken dialogue systems - Part II: Multimodal dialogue," *IEEE Transactions on Multimedia*, vol. 9, no. 3, pp. 550–566, 2007.

[24] A. Potamianos, E. Ammicht, and H.-K. Kuo, "Dialogue management in the Bell Labs communicator system," in *Proc. of Sixth International Conference on Spoken Language Processing*, Beijng, China, 2000, pp. 603–606.

[25] Q. Zhou, A. Saad, and S. Abdou, "An enhanced BLSTIP dialogue research platform," in *Proc. of Sixth International Conference on Spoken Language Processing*, Beijng, China, 2000, pp. 1061–1064.

[26] "FreeTTS," http://freetts.sourceforge.net/docs/.

[27] M. A. Walker, D. J. Litman, C. A. Kamm, and A. Abella, "PARADISE: A framework for evaluating spoken dialogue agents," in *Proc. of the Association for Computational Linguistics (ACL)*, Somerset, New Jersey, 1997, pp. 271–280.

[28] N. Beringer, U. Kartal, K. Louka, F. Schiel, and U. Türk, "Promise: A procedure for multimodal interactive system evaluation," in *Proc. of the LREC Workshop on Multimodal Resources and Multimodal Systems Evaluation*, Las Palmas, 2002, pp. 77–80.

[29] J. Sturm, B. Cranen, F. Wang, J. Terken, and I. Bakx, "Effects of prolonged use on the usability of a multimodal form-filling interface," in *Spoken Multimodal Human-Computer Dialogue in Mobile Environments*, pp. 329–348. Springer, The Netherlands, 2004.

Manolis Perakakis (StM'07) was born in Iraklion, Greece in 1974. He received the Diploma and the M.S degrees from the Dept. of Electronics and Computer Engineering, Technical University of Crete, Greece, in 2001 and 2003 respectively.

Manolis Perakakis worked as a consultant at Dialogos S.A. from 2000 to 2001. Since 2004, he is a research assistant and a Ph.D. candidate at the Dept. of ECE, Tech. Univ. of Crete.

His current research interests include distributed speech recognition, speech processing and multi-modal spoken dialogue interfaces.

Manolis Perakakis received the First Award of Excellence in Telecommunications by Ericsson for his senior year thesis. He is also a Trolltech worldwide programming contest beta application winner. He is a member of the Technical Chamber of Greece since 2004.

Alexandros Potamianos (M'92) received the Diploma in Electrical and Computer Engineering from the National Technical University of Athens, Greece in 1990. He received the M.S and Ph.D. degrees in Engineering Sciences from Harvard University, Cambridge, MA, USA in 1991 and 1995, respectively.

From 1991 to June 1993 he was a research assistant at the Harvard Robotics Lab, Harvard University. From 1993 to 1995 he was a research assistant at the Digital Signal Processing Lab at Georgia Tech. From 1995 to 1999 he was a Senior Technical Staff Member at the Speech and Image Processing Lab, AT&T Shannon Labs, Florham Park, NJ. From 1999 to 2002 he was a Technical Staff Member and Technical Supervisor at the Multimedia Communications Lab at Bell Labs, Lucent Technologies, Murray Hill, NJ. From 1999 to 2001 he was an adjunct Assistant Professor at the Department of Electrical Engineering of Columbia University, New York, NY. In the spring of 2003, he joined the Department of Electronics and Computer Engineering at the Technical University of Crete, Chania, Greece as an associate professor.

His current research interests include speech processing, analysis, synthesis and recognition, dialog and multi-modal systems, nonlinear signal processing, natural language understanding, artificial intelligence and multimodal child-computer interaction.

Prof. Potamianos has authored or co-authored over eighty papers in professional journals and conferences. He is the co-author of the paper "Creating conversational interfaces for children" that received a 2005 IEEE Signal Processing Society Best Paper Award. He holds four patents. He is a member of the IEEE Signal Processing Society since 1992 and he is currently serving his second term with the IEEE Speech Technical Committee.