Multimodal System Evaluation using Modality Efficiency and Synergy Metrics

Manolis Perakakis
Dept. of Elec. & Comp. Engineering
Technical Univ. of Crete
Chania 73100, Greece
perak@telecom.tuc.gr

Alexandros Potamianos
Dept. of Elec. & Comp. Engineering
Technical Univ. of Crete
Chania 73100, Greece
potam@telecom.tuc.gr

ABSTRACT

In this paper, we propose two new objective metrics, relative modality efficiency and multimodal synergy, that can provide valuable information and identify usability problems during the evaluation of multimodal systems. Relative modality efficiency (when compared with modality usage) can identify suboptimal use of modalities due to poor interface design or information asymmetries. Multimodal synergy measures the added value from efficiently combining multiple input modalities, and can be used as a single measure of the quality of modality fusion and fission in a multimodal system. The proposed metrics are used to evaluate two multimodal systems that combine pen/speech and mouse/keyboard modalities respectively. The results provide much insight into multimodal interface usability issues, and demonstrate how multimodal systems should adapt to maximize modalities synergy resulting in efficient, natural, and intelligent multimodal interfaces.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Evaluation/methodology; Voice I/O; Natural language; Graphical user interfaces (GUI)*

General Terms

Experimentation, Human Factors, Measurement, Performance

Keywords

Input Modality Selection, Mobile Multimodal Interfaces

1. INTRODUCTION

Evaluation [8] of multimodal interfaces is an important and complicated issue. Although some efforts [13, 1] have been proposed that attempt to build a unifying framework for the evaluation of speech and multimodal interfaces, there are various difficulties and issues in applying these methodologies [7]. Thus, in practice, evaluation of multimodal systems is based on traditional metrics used in human-computer interaction. Objective metrics such as speed, number of errors and task completion are usually computed for the various system configurations along with subjective metrics [5] and are statistically analyzed [9, 10] to determine the best system.

We believe that evaluation of multimodal systems should additionally encounter the principle of synergy. *Synergy* is a design principle that applies to systems that support more than one input or output modalities. Synergistic multimodal interface design can achieve multimodal interface performance that is better than the performance of constituent unimodal interfaces. A synergistic multimodal interface is more than the sum of its parts. Designing multimodal interfaces that effectively combine modalities [2], exploit synergies, are robust and adapt to the users, is not a trivial task. Design principles [12] such as compositionality and consistency should be applied and the resulting interfaces should be carefully evaluated in an effort to make subsequent interface improvements.

As an example of highly synergistic interface design we consider speech & GUI multimodal interfaces. It is widely accepted that speech [6] and graphical user interfaces (GUI) when combined to create a multimodal system offer high complementarity [4, 3]. As far as input is concerned, speech can be a more efficient and natural modality for entering information into a system, but it is lacking in robustness and consistency; in contrast, GUI interfaces have low error rates and offer easy error correction. As far as output is concerned, visual output is fast (parallel) compared to much slower (sequential) speech output. Thus, multimodal systems that combine GUI 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 presentation of output using the visual vs. the auditory modality, (iii) "error-correction" synergy, i.e., correcting speech recognition errors via the GUI.

Our goal in this study is to go beyond traditional objective metrics and propose new metrics that better explain how factors such as unimodal efficiency, input modality selection, interface design and exploitation of synergies affect the multimodal system performance. For this purpose we introduce

ICMI'08, October 20–22, 2008, Chania, Crete, Greece.

two new objective evaluation metrics "relative modality efficiency" and "multimodal synergy". Modality efficiency when compared with modality usage identifies suboptimal use of input or output modalities in the course of the interaction. "Multimodal synergy" expresses in a single number the percent of interface efficiency improvement compared to the average of the unimodal interface efficiency. "Multimodal synergy" can be used to identify problems in effectively combining various modalities. The proposed metrics are shown to be useful tools for identifying usability problems in multimodal systems.

The proposed evaluation metrics are put to the test for the evaluation of a multimodal travel reservation system that can handle speech, keyboard and GUI (mouse/pen) modalities. Evaluation shows that the novel metrics can provide good insight into usability and interface design issues for multimodal systems. Especially, "multimodal synergy" can serve as a single number that characterizes the efficiency gains over unimodal systems.

The organization of this paper is as follows. In section 2, the two new objective metrics "relative modality efficiency" and "multimodal synergy" are defined. In section 3, we briefly describe a travel reservation multimodal system running on PDA and desktop environments that is used as a case study. In section 4, the evaluation methodology is outlined. The evaluation results for the described multimodal systems are reported in section 5 and then further discussed in section 6. The main conclusions of this study and future work are presented in Section 7.

2. OBJECTIVE METRICS FOR MULTIMODAL SYSTEMS EVALUATION

Objective metrics are extensively used in HCI in order to evaluate the usability of a system. Common metrics used for the evaluation of both spoken dialogue and multimodal dialogue systems include task completion, time to task completion, number of turns, word and concept error rate. Such metrics can be computed per user, task or subtask. In addition, for multimodal systems, objective measures such as the usage of each modality (both in number turns and total duration) are used to gage the contribution of each modality to the usability of the system. Although these metrics are very useful for direct comparison between competing interface implementations and systems, the metrics themselves are often hard to interpret from a usability standpoint. In addition, most of the proposed objective metrics suffer from poor correlation with subjective usability metrics.

Next we define two new metrics that can help the system designer (in conjunction with the aforementioned metrics) gain a deeper insight into usability issues during multimodal interface design. The first metric, relative modality efficiency, calculates the amount of information communicated in unit time for each modality, i.e., the information bandwidth. Relative modality efficiency should correlate well with relative modality usage unless there is information asymmetry between the user and the system (see Section 5). The second metric, multimodal synergy, compares the multimodal interfaces with the "sum" of its unimodal parts and measures how "synergistic" the interface design is.

Note, that although we define next the two evaluation metrics for the case of a pen & speech multimodal dialogue system, the definitions are modality independent and can thus been used for any combination of modalities in any multimodal system.

2.1 Relative Modality Efficiency and Modality Usage

Modality efficiency is defined here to be proportional to the inverse of the time required by that modality to complete a task. Specifically, lets assume that T_s and T_g is the overall time spent using the speech and visual (GUI) modality respectively for a form-filling task using a multimodal interface. The number of fields (attributes) that are filled correctly using each modality is N_s and N_g respectively[1]. The relative efficiency of the speech modality (compared to the GUI modality) is defined as

$$E_s = \frac{\frac{N_s}{T_s}}{\frac{N_s}{T_s} + \frac{N_g}{T_g}} = \frac{N_s T_g}{N_s T_g + T_s N_g} \quad (1)$$

for a GUI and speech multimodal interface. *Thus efficiency is proportional to the number of tokens (filled fields) communicated correctly in unit time*, or else the *information bandwidth* of each modality.

Relative modality usage is defined here as the percent of time spent using this modality over the total interaction time. For example, for a speech and GUI system, the relative usage of the speech modality is defined as

$$U_s = \frac{T_s}{T_s + T_g}. \quad (2)$$

For a user that selects modalities based solely on efficiency consideration the ratio of modality efficiency to modality usage, E_s/U_s should be approximately one. This is equivalent to using each modality in proportion to its information bandwidth, i.e.,

$$\frac{E_s}{U_s} = 1 \;\Rightarrow\; T_s \sim \frac{N_s}{T_s}. \quad (3)$$

Ratios $E_s/U_s > 1$ signify underuse of the speech modality while $E_s/U_s < 1$ signify overuse (speech bias).

Alternatively, one can define relative modality usage in terms of the number of turns rather the time spent using each modality. Let us define Q_s and Q_g the number of speech and GUI turns, respectively. Then, the percent of speech usage is defined as :

$$QU_s = \frac{Q_s}{Q_s + Q_g}. \quad (4)$$

2.2 Multimodal Synergy

Next we define multimodal synergy as the percent improvement in terms of time-to-completion achieved by our multimodal system compared to a multimodal system that randomly combines the different modalities. In our example, where the visual (GUI) and speech modalities are combined, time-to-completion for the "random" system is computed as the weighted linear combination of the time-to-completion of the "Speech-Only" (speech input/speech output) and the "GUI-Only" (pen input/visual output) systems, with weights proportional to the usage of each modality in our actual multimodal system. Specifically, lets assume that D_s, D_g

[1]We define as field any attribute defined in the GUI that has a label and gets filled, thus a single field might contain variable numbers of concepts or words, e.g., "date" field.

and D_m are the time-to-completion of the "Speech-Only", "GUI-Only" and multimodal systems, and U_s and U_g are the relative usage of the speech and visual modalities in the multimodal system (normalized in [0,1] and summing to 1 as defined in the previous section). Then the time-to-completion of the multimodal system D_r that randomly selects a modality at each turn (respecting the a-priori probability of modality usage) is $D_r = U_s D_s + U_g D_g$. In general, $D_r = \sum_i U_i D_i$, where i sums over all available modalities. Modality synergy S_m for a multimodal system m is defined as:

$$S_m = \frac{D_r - D_m}{D_r} = 1 - \frac{D_m}{\sum_i U_i D_i} \quad (5)$$

where i sums over all modalities and corresponding unimodal systems.

Note that modality synergy expresses the relative improvement in terms of time-to-completion achieved by multimodal interfaces over the sum-of-its unimodal parts, thus the term *synergy*. Also note that synergy may be negative. For example, a multimodal system that combines modalities inefficiently, does not exploit synergies well or is difficult or complex to use (increased cognitive load) may have negative multimodal synergy.

An alternative definition of synergy is to compare the time to completion of the multimodal system D_m with the *average* time to completion of the corresponding unimodal systems $D_r^R = (1/N)\sum_{i=1}^{N} D_i$, i.e., use a "truly" random combination of the unimodal systems. Thus, the random-combination modality synergy S_m^R for a multimodal system m is defined as:

$$S_m^R = \frac{D_r^R - D_m}{D_r^R} = 1 - \frac{N\ D_m}{\sum_{i=1}^{N} D_i} \quad (6)$$

where N is the total number of available modalities. One can argue that this definition of synergy fully captures the efficiency gains due to the "input modality choices" of the user. Indeed in almost all practical situations the random-combination synergy will be greater than the multimodal synergy defined above.

Finally, note that although the discussion here focuses on input modality synergy, the formulas above capture also output or presentation synergies. If one wants to focus solely on input modality synergies, all unimodal systems used to compute D_r or D_r^R should share the same multimedia output interface. For multimodal dialogue systems this means that the unimodal speech input system should allow for graphical output, i.e., "visual feedback". This speech input/multimedia output system is abbreviated as "OMSI" in the experiments that follow[2].

3. MULTIMODAL SYSTEMS

A multimodal dialogue travel reservation system that runs on both desktop and PDA environments is described in this section (see [11] for a full description). The system supports keyboard, GUI (pen/mouse) and speech modalities. It can be used either in unimodal mode, using any of the three modalities (e.g. "Speech-Only" or "GUI-Only") or in multimodal mode by combining any of the above modalities (see below). Two different settings are described next. In the first setting the system runs on a PDA device combining speech and pen modalities. In the second setting the system combines mouse and keyboard modalities in a desktop computer.

In the first setting, two unimodal modes "GUI-Only" (GO) and "Speech-Only" (SO) and three multimodal modes are used. The three multimodal interaction modes are: "Click-to-Talk" (CT), where visual input is the default modality, "Open-Mike" (OM), where speech input is the default modality, and "Modality-Selection" (MS) where the system selects the modality that is most efficient for the average user at each turn. 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. 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 during OM mode.

Specifically, for CT interaction, pen is the default input; the user needs to click the "Speech Input" button (see Fig. 1) to override the default input modality and use speech input. For OM interaction, speech is the default input modality; the system is always listening and a voice-activity detection (VAD) event activates the recognizer. MS is a mix of CT and OM interaction; the system switches between the two interaction modes depending on efficiency considerations (the number of input choices available for the current context, that is attribute size, see Table 1). Speech input is faster compared to pen input when many input choices are available on the PDA; the threshold of 25 input choices was chosen based on the input mode efficiency of the stereotypical user. Another mode, "Open-Mike Speech-Input" (OMSI), allows only speech input while the system output supports 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.

Compared to the first setting, in the second setting, keyboard is used instead of speech input and mouse instead of pen input. Thus two unimodal systems are defined, namely "Keyboard-Only" ("KO", keyboard only input/GUI output) and "Mouse-Only" ("MO", mouse only input/GUI output). Also one multimodal mode is used instead of three different ones , the "Keyboard-Mouse" mode (KM) in which the user can use either mouse or keyboard input at each interaction turn (keyboard or mouse input/GUI output).

Note that in the desktop case, editable combo-boxes are used to allow for keyboard input in addition to mouse input. Otherwise, the desktop GUI interface is identical to that used for the PDA systems shown in Fig. 1. Next, we use the term GUI modality to refer to pen/mouse input for the PDA/desktop settings respectively.

4. EVALUATION METHODOLOGY

4.1 Evaluation setting

The evaluation setting is outlined next (see [11]). Five scenarios of varying complexity were used for evaluation: one/two/three-legged flight reservations and round trip flights with hotel/car reservation. The cumulative usage of attributes across all scenarios is shown in Table 1; the four most frequently used attributes are shown ordered by the number of available values in the grammar. We refer to the two attributes first listed, namely "city" and "airline", that

[2]It is experimentally verified that a significant portion of multimodal synergy is due to "visual feedback".

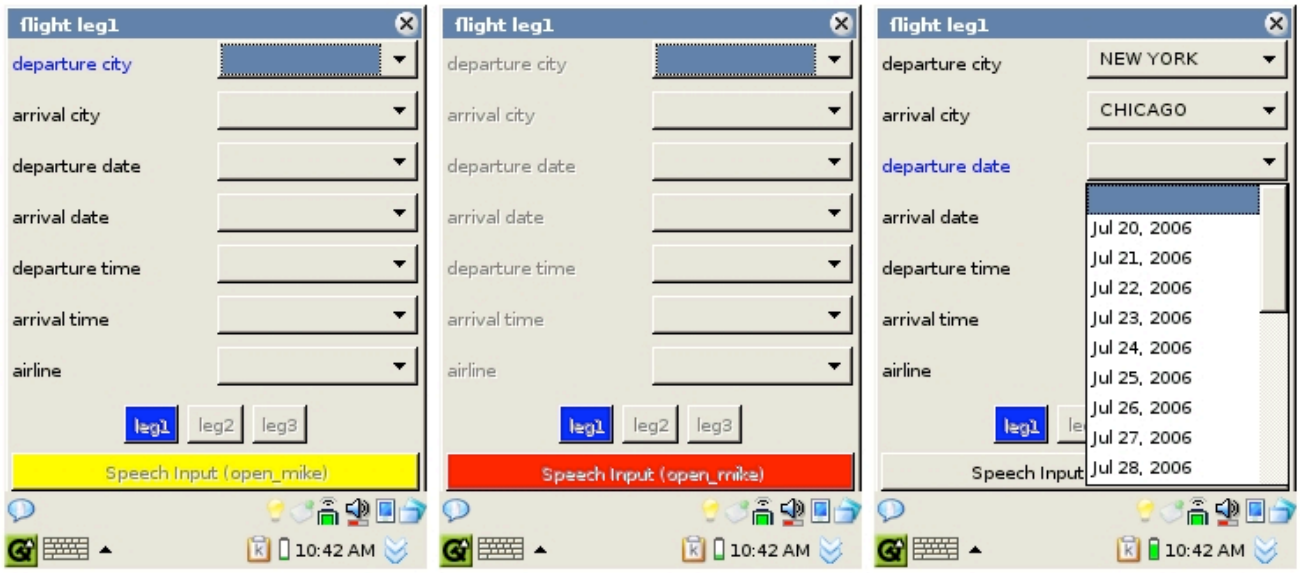

Figure 1: "Modality-Selection" interaction mode examples on the PDA. System is in "Open-Mike" mode in the first frame (speech button is yellow indicating waiting for input), receives user input "From New York to Chicago" during the second frame (speech button is red showing activity) and switches to "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 (from [11]).

have more that 25 possible values as "long" attributes while the rest ("date" and "time") are referred to as "short" attributes. Note that the cumulative attributes usage across all scenarios is about the same for "long" and "short" attributes ($14 + 5 = 19$ vs $10 + 10 = 20$). Eight non-native English-speaking users evaluated all systems on all five scenarios in random order for both desktop and PDA settings (different users for PDA/desktop systems evaluation).

Table 1: Attribute size and cumulative attribute usage across all scenarios.

attribute name	number of values	total usage
city	135	14
airline	93	5
date	22	10
time	9	10

4.2 Inactivity and Interaction Times

To better comprehend user behavior and patterns during multimodal interaction we have broken down user time into inactivity and interaction times. A schematic of the breakdown is shown in Fig 2. Inactivity time[3], 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 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 certain turn. We refer to this time as interaction time. We have found in previous studies of the same system [11], that interaction times in multimodal modes depend mainly on the input modality choice patterns (interaction time for unimodal pen/mouse input is inversely related to attribute size, see Table 1). For inactivity times, we have found that they depend on both input modality (much higher for the case of speech input) and interaction mode.

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

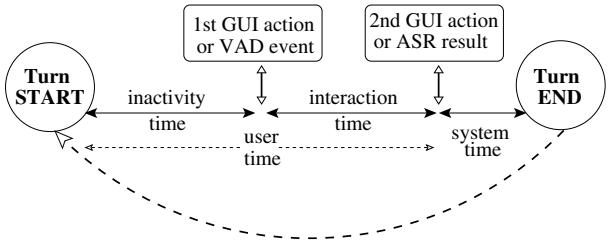

Figure 2: Turn times decomposition to user and system times. Note that user times consist of inactivity and interaction times (from [11]).

4.3 Relative Modality Efficiency & Modality Selection

Relative modality efficiency defined in Eq. 1 can be computed as a function of the interaction mode, interaction context or user, by adjusting appropriately the time T and number of tokens N in the definition. Modality efficiency results are presented for overall time, interaction and inactivity time as defined in the previous section.

Similarly relative modality usage is computed based on Eq. 2. As noted in Section 2.1, the two quantities should be

Setting	PDA speech & pen			desktop kbd & mouse
Mode	CT	OM	MS	KM
inactivity	-2.6	25.5	0.0	1.3
interaction	24.0	17.8	31.0	28.7
overall	12.7	21.1	17.8	18.0

Table 2: Multimodal synergy(%) for the four multimodal interaction modes.

plotted against each other to help us understand inefficiencies in the modality usage. By depicting the relative efficiency and modality usage in a 2D-plot for different modes, contexts and users, it is easy to identify inefficiencies that might be due to poor interface design or information asymmetries between the user and the system.

4.4 Multimodal Synergy

Likewise, synergy can be computed for each interaction context, interaction mode, user or any combination of the above, by using the appropriate time D measurements. In this evaluation, we use the random combination synergy defined in Eq. 6. Note that for the computation of multimodal synergy regarding the PDA speech & pen systems the "Speech-Only" and "GUI-Only" unimodal systems are used while for the desktop keyboard & mouse system the "Mouse-Only" and "Keyboard-Only" systems are used.

Results are derived for inactivity, interaction and overall times. The breakdown into interaction and inactivity time is especially relevant because interaction roughly corresponds to time spent on user input, while inactivity roughly corresponds to time spent on system output and cognitive processing. As a result, *interaction synergy measures input synergies*, and *inactivity synergy measures output plus cognitive load synergies*[4]. The breakdown can help the designer pinpoint usability problems in the interface design.

5. EVALUATION RESULTS

In this section the evaluation results of the two different multimodal settings (PDA speech & pen and Desktop mouse & keyboard) using the two metrics (relative modality efficiency & synergy) are presented.

5.1 PDA speech & pen multimodal systems

5.1.1 Relative Modality Efficiency

In Fig. 3(a)-(d), relative speech efficiency is plotted against relative speech usage (percent number of turns). There are three free variables in these plots, namely, interaction mode (CT, OM, MS), interaction context (city, airline, date, time) and user (u1 to u8). In all plots, a dashed line (y=x) is used to help identify efficient behavior, i.e., modality usage that is proportional to the modality efficiency. Correlation between modality efficiency and modality usage is indicated with a solid line in each plot. Note that in almost all cases, the linear regression line is located higher than the dashed line, indicating an "overuse" of the speech modality by the users, i.e., a "speech bias".

[4]Cognitive load synergy is probably a misnomer since this quantity is usually negative. This is due to the fact that the inclusion of additional input and output modalities usually increases cognitive load.

context	city(135)	airline(93)	date(22)	time(9)
PDA speech & pen multimodal systems				
inactivity	-8.1	21.6	4.9	24.9
interaction	33.1	31.5	6.6	10.3
overall	18.7	27.6	5.8	18.4
Desktop keyboard & mouse multimodal system				
inactivity	7.9	-25.0	6.0	-7.5
interaction	44.6	20.5	11.0	17.6
overall	33.0	5.1	8.4	6.0

Table 3: Multimodal synergy(%) for the four contexts (attributes).

As shown in Fig. 3(a) there are quite large differences in "relative speech efficiency" between short (time, date) and long attributes (airline, city). This is expected due to the large number of options available for long attributes in the GUI combo-box and vice-versa for short attributes. For both short and long attributes there is a clear bias towards the speech modality. Users choose speech more frequently over pen input (e.g., for the date field), despite the fact that pen input is more efficient in this case. In Fig. 3(b), results are shown for the three multimodal interaction modes. All three modes display speech bias, especially MS and CT modes, which have relative speech efficiency less than 50% and speech usage around 60%. In Fig. 3(c), the combined data points for interaction modes and contexts over all users are shown. Note that for the two long attributes (city and airline) speech usage is very high (ranging from about 80% to 90%) as expected, regardless of interaction mode. On the other hand, for short attributes (date and time) one can note that *interaction mode clearly affects input patterns*. For short attributes the data points are near the dashed line for CT and MS modes as expected, however, for OM mode, speech usage is very high (much above 70%). Thus, the default input modality (speech in this case) biases users away from efficient modality selection.

Fig. 3(d) shows the combined data points for interaction contexts and users over all modes. For long attributes, with the exception of point (city, u3) speech usage ranges between 74% and 95%. For the time attribute, with the exceptions of u3 and most notably u6, speech usage is below 50% as one would expect. For the two short attributes, only three users are GUI biased. The data points demonstrate *a "non-linear" user behavior; users abruptly switch from GUI to speech when speech becomes more efficient*. Two important observations are that: (i) the switching point is around 45% speech efficiency rather that 50% demonstrating a speech bias, and (ii) in the area of equal modality efficiencies there is high variability in modality usage demonstrating the uncertainty of the user over which modality is more efficient.

5.1.2 Multimodal Synergy

In the left part of Table 2, the synergy between the speech and GUI modality is computed for the three PDA multimodal modes. For interaction times, MS mode has the higher synergy (31%) followed by CT and then OM modes. This means that for MS mode, users selected input modality, based on unimodal efficiency consideration most of the time compared to, e.g., OM mode[5]. As far as inactivity times

[5]Recall that for OM users used speech much more often (see

are concerned, OM which by design favors speech modality choice has low inactivity times. In contrast, high use of speech in the other two modes, results high inactivity times and thus very low synergy (-2.6 for CT, 0 for MS). The *low inactivity synergy for CT and MS modes demonstrate increased cognitive load and time lost to modality switching*. Regarding overall times, synergy is higher for OM mode, followed by MS and then by CT modes. Synergy regarding overall time, can be generally thought as a weighted average of the synergies of inactivity and interaction times.

In the top part of Table 3, the synergy between the speech and GUI modality is computed for the four attributes. As far as interaction times are concerned, there is a clear separation of long and short attributes. Users exploit modality selection to use speech input in favor of pen input for the two long attributes, since as shown in Fig 3(a) the relative speech efficiency is close to 60%. In contrast to long attributes, for which synergy is above 30%, synergy for short attributes is much lower (10.3 and 6.6) since users overuse speech input despite being less efficient, compared to pen input. For inactivity times, there is high synergy for airline and time attributes but low and negative synergy for date and city attributes respectively.

In the top part of Table 4, the synergy between the speech and GUI modality is compared across the eight users. The mean and standard deviation for synergy across users is shown in the last two columns. For interaction times all synergies are positive and for some users quite high, e.g., 39% for u7. One can note high variability among the users for interaction time synergy and even higher for inactivity time synergy. Some users even show negative synergy, such as u4 and u5, demonstrating high cognitive load. Results regarding overall time synergy, show that overall, users helped by system design, can improve considerably their performance compared to unimodal systems.

5.2 Desktop keyboard & mouse multimodal system

Compared to the speech and pen multimodal systems, the keyboard and mouse multimodal system has some common characteristics but also some notable differences. Mouse input efficiency is again related to combo size as with pen input in the PDA systems and keyboard input efficiency differs only slightly among the different contexts (as with speech input in PDA systems). On the other hand, mouse input in desktop systems is generally faster compared to pen input in the PDA environment which means that difference in unimodal efficiency between mouse and keyboard is less for "KM" mode. Next we present the results for both relative modality efficiency and multimodal synergy.

5.2.1 Relative Modality Efficiency

In Fig. 3(e) & Fig. 3(f), relative relative speech efficiency is plotted against relative keyboard usage (percent number of turns) for the "KM" mode. As shown in Fig. 3(e) relative keyboard efficiency is above 55% for long attributes (city and airline) and about 45% for time attribute (for date it is close to 50%). Nevertheless it is a bit suprising that users preferred to use mouse input for the airline context (this may be attributed to the fact that most airline values were somehow long to write, e.g. northwest, southwest). As shown in Fig. 3(f), with the exception of user u8 and user u2 (the only with clearly high keyboard usage), all rest users preferred to use mouse input most of the time despite being less efficient compared to keyboard input.

discussion on speech overuse regarding Fig. 3(d)).

5.2.2 Multimodal Synergy

As shown in right part of Table 2, the synergy between the keyboard and mouse modality is high as far as interaction and overall times are concerned while it is almost zero for inactivity times. Since "visual-feedback" is the same for both "MO" and "KO" modes, there is no gain as far as inactivity times are concerned. Additionally because error rates are close to zero ("error correction" synergy zero) interaction synergy is almost due to "input modality choice" synergy alone.

In the bottom part of Table 3, the synergy between the keyboard and mouse modality is computed for the four attributes. Note that results for interaction time synergy follow a similar pattern compared to the PDA systems. Synergy regarding inactivity times is negative for time and especially for the airline attribute. As an effect overall synergy is high only for the city attribute.

In the bottom part of Table 4, the synergy between the keyboard and mouse modality is compared across the eight users. The mean and standard deviation for synergy across users is shown in the last two columns. One can note that again variability is high among the eight users, especially as far as inactivity times are concerned.

6. DISCUSSION

"Input modality choice" synergy is more clearly pronounced in the case of context results for interaction times, shown in the top part of Table 3 and Fig. 3(b), for which differences in unimodal efficiency are quite large (long attributes). This causes a clear decision on behalf of the users regarding modality choice; users almost always use speech input, except in the case of speech recognition errors for which they use GUI input. In contrast, for short attributes, relative speech efficiency is closer to the 50% decision line, thus making more blurry the modality choice decision. This also holds for the desktop case. Comparing Fig. 3(a) and Fig. 3(e) one can see that in the desktop case, difference in unimodal efficiency between keyboard and mouse modalities is small compared to the difference in unimodal efficiency between speech and pen in the PDA case. As an effect, input modality selection becomes more blurry and users select mouse input not only for the short (date & time attributes) but also for the airline attribute (see Fig. 3(e)).

All interaction synergy results are positive, highlighting that "input modality choice" (added with "error-correction" in the case of PDA speech stems) synergy is not only high for all systems but also significantly dominates the overall time synergy. With the exception of "OM" mode, inactivity time synergy is low or even negative for all modes, indicating that e.g. "visual-feedback" synergy in the case of PDA systems is counterbalanced by the high "cognitive load" imposed by these more complex multimodal systems.

Variability in interaction synergy (Table 4) is high among the users, indicating that multimodal modes may not serve equally well all users. Note that user synergy expresses the percent efficiency improvement of combined versus unimodal usage for a certain user. This means for example that user u7 (interaction synergy 39%) during multimodal interaction, exploited input modality and other synergies in

User	u1	u2	u3	u4	u5	u6	u7	u8	mean	std
PDA speech & pen multimodal systems										
inactivity	16.4	21.4	8.4	-21.1	-2.7	9.6	24.8	2.5	7.4	14.7
interaction	26.5	33.2	15.5	30.5	17.2	14.4	39.0	13.4	23.7	9.85
overall	22.8	28.2	12.5	11.0	10.0	12.0	32.5	8.2	17.2	9.33
Desktop keyboard & mouse multimodal system										
inactivity	14.4	10.0	0.7	-24.4	5.1	-9.0	12.3	12.0	1.3	13.6
interaction	33.6	33.0	41.0	26.4	23.5	20.9	28.0	20.0	28.7	7.2
overall	25.8	24.6	25.7	7.2	13.9	7.7	22.0	16.5	18.0	7.7

Table 4: Multimodal synergy(%) for the eight users

a higher degree, that helped him improve his performance with the system more, compared to other users. The differences in synergy are due to user dependent input modality efficiency and usage, variable speech recognition rates, and, most-importantly, to what degree users used efficiency considerations when selecting the input modality at each part of the interaction[6]. In any case, the fact that synergy is highly user-dependent shows that there is potentially high-reward in designing multimodal interfaces that *adapt to the user*. Creating multimodal interfaces that are "optimal" for a stereotypical user does not grep all the reward (in terms of synergy) over unimodal interfaces.

7. CONCLUSIONS

In this paper, metrics for measuring "multimodal synergy" and relating modality usage with unimodal efficiency, were devised and evaluated with two different multimodal systems. Both metrics showed their utility and generalizability across modalities and systems and provided much insight into multimodal interface design. Relative modality efficiency, showed that although on average, users tend to use the most efficient modality at each turn, modality usage patterns were highly user dependent and that the design of the multimodal interface can affect user behavior e.g. excessive use of speech in "Open-Mike" mode. In addition, "multimodal synergy" was shown to be a valuable tool for investigating usability problems related to modality fusion and fission at the interface level, as well as usability problems associated with cognitive load.

Based on these observations we intend to use these metrics in the evaluation of alternative multimodal systems that use a variety of modalities, in order to ensure their utility and generalizability across systems and modalities. Future work will also focus on how to exploit evaluation results based on these metrics to design adaptive and more efficient multimodal interfaces.

Acknowledgments 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%).

8. REFERENCES

[1] N. Beringer, U. Kartal, K. Louka, F. Schiel, and U. Turk. PROMISE: A Procedure for Multimodal Interactive System Evaluation. *Proceedings of the Workshop on Multimodal Resources and Multimodal Systems Evaluation*, pages 90–95, 2002.

[2] V. Bilici, E. Krahmer, S. t. Riele, and R. Veldhuis. Preferred Modalities in Dialogue Systems. *Sixth International Conference on Spoken Language Processing*, pages 727–730, 2000.

[3] P. Cohen, M. Johnston, D. McGee, S. Oviatt, J. Clow, and I. Smith. The Efficiency of Multimodal Interaction: a Case Study. *Fifth International Conference on Spoken Language Processing*, 1998.

[4] M. Grasso, D. Ebert, and T. Finin. The Integrality of Speech in Multimodal Interfaces. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 5(4):303–325, 1998.

[5] K. Hone and R. Graham. Towards a tool for the Subjective Assessment of Speech System Interfaces (SASSI). *Natural Language Engineering*, 6(3 & 4):287–303, 2001.

[6] J. Lai and N. Yankelovich. Conversational speech interfaces. In *The human-computer interaction handbook: fundamentals, evolving technologies and emerging applications*, pages 698–713. Lawrence Erlbaum Associates, Inc., Mahwah, NJ, USA, 2003.

[7] L. Larsen. Issues in the evaluation of spoken dialogue systems using objective and subjective measures. *IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 209–214, 2003.

[8] D. Litman and S. Pan. Designing and Evaluating an Adaptive Spoken Dialogue System. *User Modeling and User-Adapted Interaction*, 12(2):111–137, 2002.

[9] R. Mason, R. Gunst, and J. Hess. *Statistical Design and Analysis of Experiments*. Wiley, 1989.

[10] J. Myers and A. Well. *Research Design and Statistical Analysis*. Lawrence Erlbaum Associates, 2003.

[11] M. Perakakis and A. Potamianos. A study in efficiency and modality usage in multimodal form filling systems. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(6):1194–1206, Aug. 2008.

[12] L. Reeves, J. Martin, M. McTear, T. Raman, K. Stanney, H. Su, Q. Wang, J. . Lai, J. Larson, S. Oviatt, et al. Guidelines for multimodal user interface design. *Communications of the ACM*, 47(1):57–59, 2004.

[13] M. Walker, C. Kamm, and D. Litman. Towards developing general models of usability with PARADISE. *Natural Language Engineering*, 6(3 & 4):363–377, 2001.

[6]This last factor is directly related to synergy, random input modality selection achieves zero synergy.

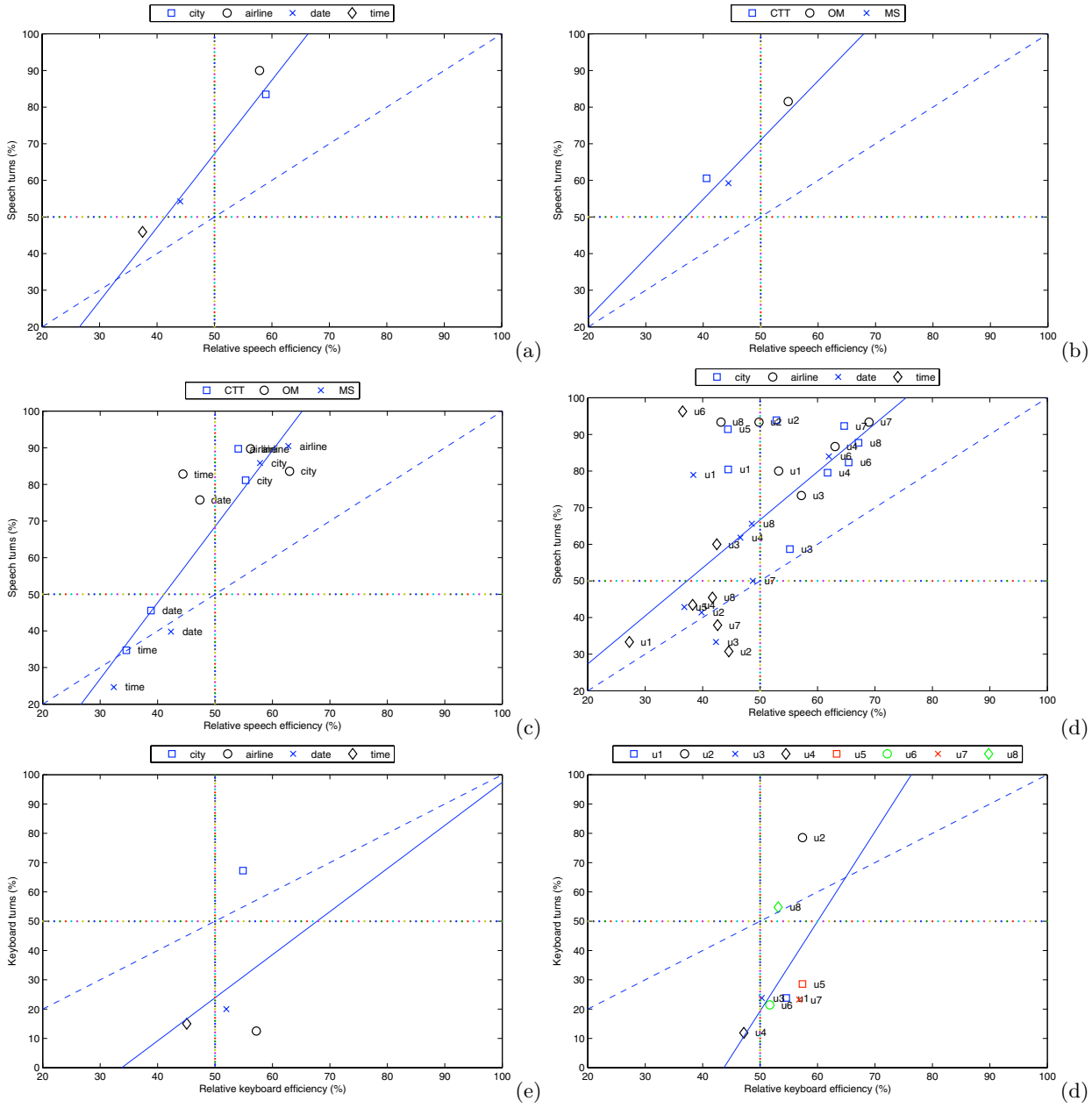

Figure 3: Modality usage (speech for plots (a)-(d) and keyboard for plots (e)-(f)) as a function of relative modality efficiency - overall times are shown. (a) context averaged over users and interaction modes (4 points). (b) interaction mode averaged over users and contexts (3 points). (c) combined data points for interaction modes and contexts over users (12 points). (d) combined data points for users and context over interaction modes (32 points). (e) context averaged over users (4 points). (f) user averaged over contexts (8 points).