Text versus Speech: A Comparison of Tagging Input Modalities for Camera Phones

Mauro Cherubini, Xavier Anguera, Nuria Oliver, and Rodrigo de Oliveira
Telefónica Research
via Augusta, 177 – 08021 Barcelona, Spain
{mauro, xanguera, nuriao, oliveira}@tid.es

ABSTRACT

Speech and typed text are two common input modalities for mobile phones. However, little research has compared them in their ability to support annotation and retrieval of digital pictures on mobile devices. In this paper, we report the results of a month-long field study in which participants took pictures with their camera phones and had the choice of adding annotations using speech, typed text, or both. Subsequently, the same subjects participated in a controlled experiment where they were asked to retrieve images based on annotations as well as retrieve annotations based on images in order to study the ability of each modality to effectively support users' recall of the previously captured pictures. Results demonstrate that each modality has advantages and shortcomings for the production of tags and retrieval of pictures. Several guidelines are suggested when designing tagging applications for portable devices.

Keywords

Photo tagging, text tagging, audio tagging, camera phones, personal image search

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: Miscellaneous—*input devices and strategies, interaction styles, voice I/O*

1. INTRODUCTION

Let us consider the following scenario: *Paul meets Mary in a restaurant downtown. While they catch-up on the recent events of their lives, Mary mentions that she met John –a friend they have in common– a couple of weeks earlier during a birthday party. She took a picture of John and his fiancé during the party. Paul is eager to see the picture as he has not seen John for more than five years. Mary starts scrolling the list of filenames of the pictures stored in her phone ...*

Most of the mobile phones currently available today on the market include a camera. Users are increasingly taking advantage of the ubiquity of their camera phones to capture, share and archive mementos of their lives. Moreover, this sharing often happens when people are face-to-face [11]. The increasing storage and computing capabilities of mobile devices open up the challenge of *mobile multimedia information retrieval*: tools for assisting users in retrieving the *right* multimedia content at the *right* time while on-the-go, as suggested by the previously described scenario.

There are still a number of open questions in this area. While some propose strategies for *automatic* indexing of pictures [25, 21], others focus on lowering the barriers for *manual* annotation [13]. Our work is related to the latter, but with a different perspective: assuming[1] that users are willing to input at least one tag per picture right after the image is captured, we investigated which input modality, namely *text* via the keypad, and *speech* through the microphone, is more effective in supporting annotation and retrieval of pictures on a mobile device.

As multimodal mobile interfaces become more pervasive, comparative studies of each of the input modalities –and their combinations– in the context of a specific *task* are necessary to better understand when it makes sense to use one or another. The contribution of this paper is to offer qualitative and quantitative comparisons of *text vs speech* in the context of a photo annotation and retrieval task on a mobile phone. In particular, the following research questions are addressed:

1. How does the availability of different tagging modalities –text, speech or both– influence the production of tags and the retrieval of pictures on mobile devices?
2. What are the major design implications?

In this paper, results are reported of a month-long field study during which 20 participants collected and annotated pictures with their mobile phones. They were assigned to different experimental groups in which they could annotate their pictures using either text, speech, or both (see Section 4.5). The same participants took part in a controlled experiment, where they were asked to retrieve images using the annotations they had created (see Section 4.1). This procedure helped in understanding which input modality is more effective to support users' recall in the given context, thus highlighting their advantages and shortcomings for the production of tags and retrieval of pictures (see Section 6). From these findings, a set of guidelines are proposed for the design of mobile applications targeting photo retrieval via multimodal annotations (see Section 6.1).

2. RELATED WORK

We shall summarize next the most relevant previous work in the areas of multimodal photo annotation and retrieval,

[1] We created this assumption because the presence of one tag per picture –at least– was necessary to grant retrieval of the pictures in the controlled experiment as explained in Section 4.1. It is out of the scope of this study to prove that users are willing to tag every picture, a few or none of them.

MobileHCI'09 September 15-18, 2009, Bonn, Germany
ACM 978-1-60558-281-8.

both in PCs and mobile phones.

Speech and Text Photo Annotations.

In the last decade, many prototypes have been designed both in industry and academia to support the annotation of digital pictures on desktop computers. For example, Tan and colleagues [24] developed **SmartAlbum**, a multimodal photo annotation system that unifies two indexing approaches: content-based and speech. The application allows users to search for specific features of the pictures, such as the presence of sky or water, and to use speech annotations. In SmartAlbum, audio annotations are transcribed to text and then used for indexing the picture. Similarly, Chen et al. [6] proposed the use of structural speech –as opposed to free form speech– syntax to annotate photographs in four different fields, namely event, location, people, and date/time. This was proven to further boost the performance of the speech recognition engine. Note that current state of the art in speech recognition is still far from being error-free, especially when dealing with unconstrained speech input.

Other research prototypes have exploited voice annotations, such as the **Show&Tell** system [22] where voice is combined with automatic indexing techniques, and the **FotoFile** prototype [12] for generic multimedia objects (*e.g.* short texts, videos, etc.). It is also worth mentioning the work by Stent and Loui [23], who proposed a combination of speech and text-based annotations to improve the indexing of consumer photographs. Finally, Rodden and Wood [19] carried out a longitudinal research study on how people manage their collections of digital photographs on their desktop computers. The authors asked 13 subjects to use a research prototype named **Shoebox** [15] during 6 months to catalogue their pictures using text and/or voice. The authors found that Shoebox' tagging capability was not used because participants relied on their memory and on the temporal sequence of the pictures to retrieve them. Therefore, Shoebox did not help them to increase their retrieval efficiency.

Most of these approaches involve CPU-intensive algorithms that are not feasible to port to mobile platforms given their current characteristics. Addressing this problem, Hazen *et al.* [10] recently proposed a client-server based mobile phone application that allows users to use speech to annotate and retrieve digital photographs. Their system is implemented as a light-weight mobile client connected to a server that stores the digital images and their audio annotations. The server hosts a speech recognizer for parsing input and metadata queries. Preliminary experiments demonstrated successful retrieval of photographs using *purely* speech-based annotations.

While speech-based mobile applications are still in their infancy, text-based annotation systems seem to have been better suited for mobile interfaces. In this area, several commercial products are found. **ShoZu**[2] is a mobile phone application that allows the user to post pictures, videos, or short textual messages to different commercial online multimedia repositories while tagging these media with textual tags. Similarly, the **ZoneTag** [1] application allows users to take pictures with camera phones and assign textual tags to the photos. Additionally, it uses a location-based component to suggest tags created by other users who took pictures in the same area.

While the previously described prototypes have provided evidence of the benefits of integrating different annotation modalities to the indexing of photographs, it is important to review work from other scholars who have studied how capturing and tagging pictures on a mobile phone might differ from other forms of photography.

Qualitative Studies of Picture Tagging on Mobiles.

Ames and Namaan [1] conducted a comprehensive study of people using the ZoneTag system previously described. The authors defined eight intertwined reasons of why people use tags. These could be described by the different functions –*e.g.*, organization or communication– or social goals –*e.g.*, directed to self or to others– of the pictures.

Similarly, Kindberg et al. [11] conducted research on how people use camera phones. The authors interviewed 34 subjects in an effort to understand what they photographed and why. The authors found that people took pictures with their mobile phones for affective reasons: to enrich a mutual experience by sharing an image with those who were present at the time of capture. The authors highlighted that the ability to spontaneously show images was the key value of the camera phone. They suggested that easily finding and browsing images on the phone should be made possible. The interviewed participants reacted positively to the idea of enriching this content with contextual information. Finally, they found that image browsing when many pictures were present on the mobile phones was ineffective for image retrieval.

Next, we present a comparison of different input modalities in their ability to support picture tagging and retrieval. Unfortunately, little research has been conducted to date on comparing speech and text as annotation mechanisms in general, and in mobile devices in particular.

Quantitative Comparisons of Speech and Text as Modalities for Input and Retrieval.

PC Input. Hauptmann and Rudnicky [9] conducted a controlled user study where participants had to enter a set of alphanumeric strings in the computer. Subjects either used their voice, the keyboard or a combination of the two. Utterance accuracy results showed that subjects using *speech as input required more time* to complete the task (*e.g.*, enter a number correctly) than those who where typing. Additionally, the authors cautioned how real world tasks that typically require more keystrokes per syllable, would better illustrate the effectiveness of speech. The authors concluded that speech could offer advantages for casual users, depending on the task. The more a task would require visual monitoring of the input, the more preferable speech would be as an input modality.

Similar results have been obtained by Hah and Ahlstrom [8], who ran an experiment that compared an automatic speech recognition system with keyboard and mouse as text input modalities on PCs. They computed execution time by factoring in correction time for both conditions. They showed that participants took significantly *longer in the speech* than in the typing condition. Regardless of whether participants were fast or slow typists, all of the participants preferred typing to speaking and performed better in the typing condition than in the speech condition.

A somewhat reversed relation was demonstrated by Mitchard and Winkles [16], who conducted a comparison of input modalities for data entry tasks in the context of military messages. They reported that, in the case of short messages,

[2] `http://www.shozu.com/`, last retrieved January 2009.

speech only competes with keyboard and mouse if the typist's speed is below 45 words per minute[3].

In synthesis, and given the state-of-the-art on speech recognition, previous work has shown that keyboard and mouse are more efficient modalities than speech for entering information in a desktop environment. However, it is not clear that these results are directly applicable to mobile phones. In particular, the constrained keypads on mobile devices might slow down the typing speed. In addition, one might expect an advantage of speech over typed text when the user's hands are busy, as it is often the case in mobile scenarios.

Mobile Input. Few comparisons of these modalities have been conducted on mobile devices. Perakakis and Potamianos [18] evaluated the form-filling part of a PDA multimodal dialogue system. Their results showed that *multimodal systems outperform unimodal systems* in terms of objective and subjective measures. In addition, the authors found that users tended to use the most efficient modality, depending on the situation and that there was a bias towards the speech modality. Another relevant piece of research was conducted by Cox and colleagues [7], who compared speech recognition with multi-tap and predictive text entry in the context of SMS message input on a mobile phone. Their results showed that speech is faster that the other two methods and that a combination of input methods provide the quickest task completion times. Further analysis confirmed that participants were willing to trade accuracy for speed. Precisely, this last result might distinguish the task of writing an SMS from that of tagging. While humans (recipients of the SMS) are good at inferring sense from mistyped words, most retrieval systems might produce low accuracy results with tags containing typos or slangs. Similarly, the study presented in this paper is different from a large body of work that compares input speech and typed text for menu navigation (see for instance the work of Lee and Lai [14]). Although related, the state-of-the-art in speech-based menu navigation has high performance standards, still far from those achieved by the best systems dealing with unconstrained speech input.

PC Retrieval. With respect to multimodal retrieval, Rudnicky evaluated in a subsequent study text and speech as input modalities in a data-retrieval task on a desktop computer [20]. Rudnicky's work shows that users *prefer speech-based queries* despite its inadequacies in terms of classic measures of performance, such as time-to-completion. Similar results were obtained by Mills and colleagues [15], who conducted a retrieval experiment over a small collection of pictures collected with Shoebox. The retrieval task consisted of locating photographs by their associated typed and transcribed annotations. Precision of retrieval with *typed keywords over-performed* retrieval with speech queries because of inaccuracies introduced by the speech recognition software. Brown and colleagues [5] also noticed the same inaccuracy problems. Interestingly, while users perceived speech queries as effective, objective measures of retrieval performance denied an advantage of speech over text.

Mobile Retrieval. Little research was found comparing typed text and speech in their ability to support picture retrieval on a mobile device. One of the most relevant contributions in this area is the work of Paek and colleagues [17]. They designed and evaluated **Seach Vox**, a mobile search interface that allows the user to search by means of speech input and to assist the recognizer via text hints. The multimodal interface was proven to improve search accuracy. However, the authors did not compare it with each of the modalities by themselves.

The study presented in this paper focuses on comparing typed text, speech and their integration in their ability to support annotation and retrieval of photographs on mobile phones. Two challenges shall be highlighted in the context of this work: First, it is not clear whether the results of previous work –which was mostly conducted on desktop computers– might hold for mobile devices; second, in some occasions the comparisons of these modalities conducted on mobile devices by different researchers point to opposite results. For instance, text-based annotations have been shown to improve the retrieval process when compared to spoken material which might be inaccurately transcribed [5, 15]. However, speech has been suggested to be more efficient than text for operating in a mobile device, because it gives users more freedom in terms of its input and it enables fast descriptions of complex properties [7, 10, 18].

2.1 Hypotheses

The hypotheses of our user study summarize four of the major findings in the previous work literature related to image tagging and retrieval on mobile phones:

H1: Speech is preferred to text as an annotation mechanism on mobile phones (objective measure). Quantitative comparisons of typed text and voice on PCs demonstrated an advantage of text as an input modality [9, 8]. This relation was proven to be reversed when the user is a slow typist [16] and if the task requires visual monitoring [7]. The slower input on a mobile keypad –when compared to a standard keyboard– and the possibly limited attention of a user while s/he is on the move leads to this first hypothesis.

H1-bis: Speech annotations are preferred by users even if this means spending more time on the task (subjective measure). Looking at the user's preferences, Perakakis and Potamianos [18] demonstrated a preference bias for speech as an input modality in PDAs. Similarly, this hypothesis addresses H1 in a subjective manner.

H2: The longer the tag, the larger the advantage of voice over text for annotating pictures on mobile phones. The study of Hauptmann and Rudnicky [9] concluded that voice would be faster than text as an input modality when having to enter multiple words on a mobile phone. The length of a text tag should correlate with the time required to input the tag correctly.

H3: Retrieving pictures on mobile phones with speech annotations is not faster than with text (objective measure): The work of Rudnicky [20] and Mills et al. [15] compared speech with other input modalities in desktop PCs and failed to demonstrate an advantage of speech using objective measures. Hypothesis H3 extends this conclusion to mobile devices.

3. THE MAMI PROTOTYPE

The user study presented in this paper employed a mobile phone application, named **MAMI** (*i.e.* Multimodal Automatic Mobile Indexing) [2], to add multimodal metadata to photographs at the time of capture. Figure 1 illustrates four interfaces available in MAMI. When the user takes a picture with MAMI, s/he can add one or more speech and/or

[3] People naturally speak faster than they type. An experience typist can reach approx. 80 wpm, while normal speech reaches approx. 200 wpm [7].

textual tags by means of the *indexing interface*, shown in Figure 1a. These tags are associated with the image and they are indexed in a local database. Upon indexing, the system computes and stores acoustic and image descriptors and additional metadata information such as: the textual tag, location, date and time of capture, and user ID.

At a later time, when the user desires to search for and retrieve a specific image, s/he can search the system via a speech or textual query by means of the *search interface* (Figure 1b). In the case of speech, the query is processed by MAMI to compute the query's descriptor that is compared to all other descriptors in the local database. In the case of text, the input string is compared with all other textual tags in the database. The 9 pictures that best match –*i.e.* whose descriptors are the closest to– the user's query are then retrieved, as depicted in Figure 1c. MAMI's speech matching algorithms are comparable in their performance to state-of-the art speech recognition systems due to its robustness against noise. We direct the reader to [3] for a detailed description of the MAMI prototype.

The MAMI prototype has the following characteristics:

1. *All processing is done on the mobile phone*, so that the system is always functional, independently of roaming or connectivity.
2. *Speech is not converted to text*, in order to avoid heavy computational requirements typically associated with speech recognition. In addition, MAMI's speech processing is speech independent.
3. *MAMI uses Dynamic Time Warping* (DTW) in order to compare acoustic descriptors, and the *edit distance* to compare textual tags.

In the context of the study, users were requested to enter at least one tag per picture. This was necessary for the controlled retrieval experiment as described below.

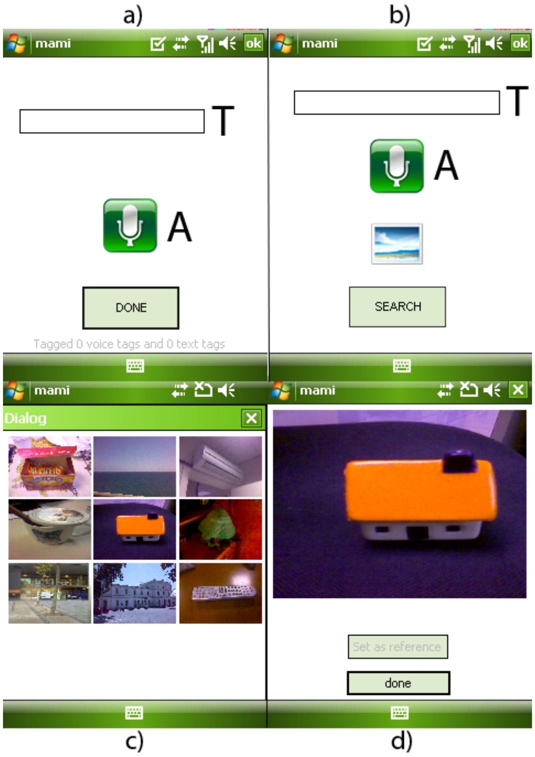

Figure 1: Screenshots of the MAMI interface: a) indexing; b) search; c) query results; and d) result preview. The text field, marked with T, was visible only in the Text and Both condition, while the audio record button, marked with an A, was visible only in the Audio and Both condition

4. METHODOLOGY

A user study was carried out in order to answer the research questions and to test the hypotheses. The study consisted of a combination of a field deployment and a controlled experiment, in a similar way to the methodology used by Hazen *et al.* [10]. This approach was selected for two reasons: First, tagging is a personal activity that evolves over time and that has an impact on different cognitive functions, such as the ability to remember the assigned tag(s) to the multimedia content. Therefore, it requires a study spanning several weeks. Second, we wanted to compare different tagging modalities not only in their ability to support image retrieval but also in how they could affect the production of tagged pictures.

4.1 Procedure

20 participants were recruited among the employees –and their families– of a large telecommunications company headquartered in Spain. Participants were first invited to an introductory session where they were assigned a mobile device with the MAMI prototype installed. During the introductory session, participants filled out a pre-study questionnaire, received an explanation of the experiment, were given basic training on how to use the MAMI prototype and moved their SIM card and contacts to the new device (so that it could be used as their main phone). Once participants were familiar with the phone and the MAMI prototype, they went away for the month-long field study.

Field Study. The deployment of the field experiment started at the beginning of August 2008, when most of the participants took their summer holidays, taking the device with them.

The field experiment lasted for 31 days (the entire month of August). During the study, participants were asked weekly and via SMS to report how many pictures they had taken so far. They were also given support in case of questions or software failures in MAMI. At the end of the field study, they were asked to participate in a controlled experiment (explained next) and to fill out a post-study questionnaire.

Controlled Experiment. The controlled study consisted of 4 image retrieval tasks where retrieval time was the main dependent variable. Participants performed three trials of the first three tasks and a single trial of the last task. Participants carried out all the tasks using the MAMI prototype on the *same* mobile phone that they had used during the field study. The tasks were designed to challenge different elements of the participants' memories, as detailed below.

The stimuli for the retrieval (*i.e.*, a picture, the playback of a speech tag or a textual description of the picture to be retrieved) were presented on the screen of a desktop computer.

T1- First task: remember the tag. It consisted of retrieving a specific picture that had been randomly selected from the pool of pictures that each participant had captured during the field study. Therefore, T1 required participants to remember the tags associated with a specific picture.

T2- Second task: remember the context. Two researchers were given three random pictures from the subject's collection and were asked to independently write six nouns describing the content of each of the pictures. For each picture, two or three nouns were selected from the intersection of the two sets provided by the researchers. For instance, the picture represented in part d) of Figure 1 might have been described by 1st researcher as "*toy, house, souvenir, games, orange, kids*", and by 2nd researcher as "*orange, miniature, baby, toy, house, furniture*". Finally selecting: "*toy, house, orange*". T2 required participants to remember the context in which the picture was taken (*i.e.*, both the picture and its associated tags), and to draw a semantic inference between the list of presented words and the tags associated with the requested picture.

T3- Third task: remember the picture. Participants were shown/played a random tag from the tags that they had assigned to their pictures. Furthermore, the tag was translated from its original modality to the other modality by one of the researchers: if the tag was textual, it was then spoken and recorded and viceversa (*i.e.* if it was a voice tag then it was typed as text). Subjects were asked to retrieve the picture that was tagged with that particular tag. This task challenged participants to remember the picture associated with a particular tag and it introduced an artificial noise comparable to that generated by speech-to-text and text-to-speech systems.

T4- Fourth task: remember the temporal sequence. It consisted of retrieving three pictures among the ten oldest and three pictures among the ten newest in the subject's collection. This task requested participants to remember the sequence of tagged events (*i.e.*, long term *vs.* short term memory).

The controlled experiment lasted for about forty-five minutes. Before starting the experiment, each participant received extensive explanations on each of the tasks, and they could participate in a training session in order to get acquainted with the tasks. Participants typically felt comfortable with the tasks after about 10 minutes of training. Special emphasis was put on showing participants how to stop the timer after they had found the requested image.

4.2 Participants

20 participants were recruited (12 male, 8 female) by email advertisement to different divisions of the company. Their mother tongue was Spanish. About a third of the volunteers were family members of employees. The median age of the participants was 30 years (min 23, max 46). Their occupations included engineers, security assistants, administrative assistants, researchers, university students, managers and graphic designers. All participants owned a mobile phone and 17 owned a digital camera and a camera phone. They were all computer literate. The pre-study questionnaire included –in addition to standard demographic questions– questions about the tagging and retrieval practices of the experimental group.

More than half of our participants reported regularly taking pictures with their mobile phone (58.8%). When asked about their favorite strategy for *organizing pictures*, 9 participants (45%) reported keeping the images in separate folders –with a folder for each event. Six participants (30%) reported using a specific software to archive their pictures, while 3 participants (15%) kept their pictures in a single folder, using their associated timestamp to keep them organized. Finally, 2 participants (10%) declared printing their pictures and organizing them in shoe boxes. Of those that reported using either a dedicated software or nested folders, 7 participants (35%) declared adding textual descriptions to their pictures (see Figure 2).

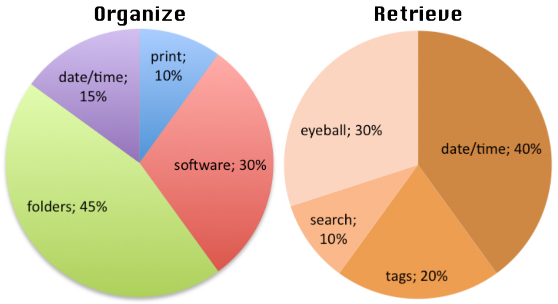

Figure 2: Favorite strategies used by participants for organizing their pictures (left) and retrieving a particular picture or a set of pictures (right)

Four participants (20%) reported searching for a specific picture once per week, 7 respondents (35%) once per month and the rest of respondents (45%) did it less frequently. Regarding the strategy that they used to *retrieve pictures*, the majority of respondents reported using the time of capture as their cue (8 or 40%) and browsing the picture thumbnails (6 or 30%). Interestingly, only 4 respondents (20%) reported using the annotations that they had previously created. The last 2 participants (10%) reported using the search functionality of the software they used to organize their pictures.

In summary, 70% of the experimental group reported having an archival and retrieval strategy that did not exploit the use of metadata to facilitate the organization and search of pictures in the collection. Additionally, only a third of the group reported using textual tags, while none used voice tags. The results of the questionnaire are consistent with the qualitative studies of picture taking presented in Section 2.

4.3 Apparatus

All participants were given an HTC P3300 mobile phone running Windows Mobile 6.0. Note that this phone has a touch screen that allows textual input via a stylus pen and small on-screen keyboard that appears in the lower half of the screen. The phone had the MAMI prototype installed. During the field study, participants were asked to use the MAMI prototype for capturing and annotating pictures on the phone.

Participants were divided in 3 experimental groups (Voice, Text, and Both), as detailed in Section 4.5. Depending on the experimental conditions, participants were allowed to see part or all the elements of these interfaces: Participants in the Text condition could interact only with the text tag field and the textual search field (UI elements marked with a T in Figure 1); conversely, participants in the Audio condition could interact only with the voice tag record button and the voice query record button (UI elements marked with an A in Figure 1); finally, participants in the Both condition could interact with all the elements of the interfaces depicted in Figure 1. In other words, participants in the Both condition could tag pictures with either text or voice –or both– at their choice.

With respect to the controlled study, participants had to retrieve the pictures given as stimuli using the MAMI on the same mobile device. The controlled study also used a desktop computer to guide users through the 4 retrieval tasks. The pictures that participants took with their phones were copied to the desktop computer, where randomly selected pictures and annotations were presented to each participant. Stimuli for each of the tasks were presented via a 19" LCD display. Advancing through the tasks required the subject to press the space bar of the keyboard connected to the desktop computer. The space bar was also used to stop the timer that registered the completion times for each trial. The PC recorded each of the participant's actions into a log file with accurate timestamps.

In addition, the mobile phone had been instrumented in order to record the users' actions with accurate timestamps. The clock of the mobile device and the PC were synchronized at the beginning of each experiment. After each experiment, the log file on the mobile phone was transferred to the main desktop computer.

Finally, a third log file contained all the actions that participants had performed during the field trial. These log files were automatically parsed to extract performance metrics and process variables as described next.

4.4 Measures

Several process variables were measured in the field study: (a) The number of sessions and the session length, where a session starts when launching the MAMI application and ends upon exiting MAMI; (b) the editing time required for the user to assign the tags to a picture; and (c) the total number of pictures taken, the average number of tags per picture, and the temporal distribution of pictures and tags over time.

During the controlled experiment, we measured: (a) The number of queries issued by the participants during each trial; and (b) the number of pictures previewed before the best match was chosen and the timer stopped. Subjects could run as many queries as needed and stop whenever they felt there was nothing else they could do in order to retrieve the given picture. If participants did not find the requested picture, the trial was flagged as a *retrieval error.*

The main performance variable that was computed during the controlled study is the time required to complete each trial. This is the time elapsed between stimulus presentation (*i.e.* showing a specific picture to retrieve) and the moment when the participant pressed the space bar to stop the timer. As voice queries took in average 6 seconds longer than text queries (due to processing time) to be executed, the time the subject had to wait for each query (text or voice) to return its results was removed from the trial completion time.

Before stopping the timer, participants were asked to preview in full screen (on the phone) the image that they considered correct. The average completion time was computed for each task using the data from all the trials, except when the user could not retrieve a given picture. In this latter case, the trial was flagged as *retrieval error* and the trial time was not used for computing the average task time. A secondary measure of performance was given by the number of pictures wrongly identified as corresponding to the stimulus, which will be referred to as *false positives.* False positives were computed for each task.

4.5 Independent Variable

Participants were randomly assigned to three experimental groups. As explained before, the MAMI prototype was customized in order to reflect the experimental conditions to which they belonged.

Seven participants were assigned to the VOICE condition, where they could tag and retrieve only with speech tags. Seven other participants were assigned to the TEXT condition, where they could tag and retrieve only with textual tags. Finally, the remaining six participants were assigned to the BOTH condition, where they could tag and retrieve pictures with either text or voice at their choice. Participants were gender balanced in each group in order to avoid gender biases. The design was therefore a standard single factorial design, where the availability of a tagging modality (VOICE vs. TEXT vs. BOTH) was a between-subjects factor.

The sample was unpaired and organized in 3 treatments. All the collected measures were expressed in interval scales. Before testing our hypotheses, we verified the assumptions of homosedasticity and normality of distributions in the experimental groups. ANOVA was therefore conducted with a Bonferroni post-hoc test[4].

5. RESULTS

The results presented in this section correspond to analyzing the data of all participants but three which encountered a small glitch in the MAMI prototype (about 25% of their data did not get properly stored during the field study). Therefore, it was decided to exclude them from the analysis. This choice did not impact the gender balance nor the balance on the number of participants in each group (*i.e.*, VOICE: 6, TEXT: 6, BOTH: 5).

5.1 Field Study

During the field trial, participants collected a total of 6279 pictures (median 205 pictures), and a total of 9741 tags (median audio 55, median text 101). They interacted with MAMI in 84 sessions (median) with an average median length of 257.8 seconds per session. Figure 3 presents a timeline visualization of the number of pictures –normalized by total number– taken in each condition for each day of the experiment. The plot illustrates how participants slowly reduced their activity until the third week and then increased it right before returning the phone. The log files of the field study provided relevant findings for H1 and H2. Concerning **H1**, it was found that participants in the BOTH condition tagged slightly more with text than with audio. However, this difference was not found to be significant. A paired t-test was conducted between the total number of textual and audio tags assigned by each participant in the different groups (mean [std] – Voice: 49.40 [40.16] *vs.* Text: 150.00 [193.46] tags, paired t(5) = -.54, p>0.05, ns). We also aggregated this measures at a picture level and counted the number of times a subject preferred only Voice, only Text, or Both tagging modalities, regardless of the number of tags. Results are reported in the table 1. Three participants tagged preferably with Text, 1 with Voice, and the last one with Both modalities. Also, this choice is not related with the number of pictures taken by the subject. Therefore, these results contradict the first hypothesis, which was predicting a preference for the speech annotation modality.

The analysis of the questionnaires that were filled after the controlled experiment provided relevant results for the second hypothesis, **H1-bis**. All participants who were assigned to the BOTH condition ($N = 6$) reported preferring text as a

[4] With the exception of the results of the questionnaire. The low number of observations did not allow hypothesis testing. These results are therefore purely speculative.

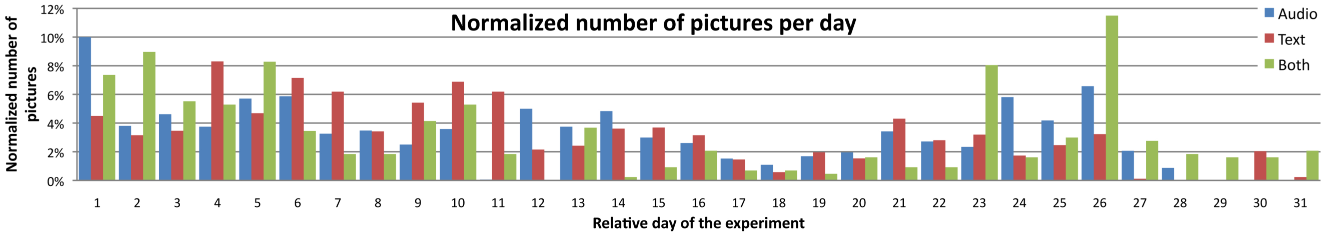

Figure 3: Timeline visualization of the number of pictures produced in each day of the experiment

Table 1: Cross-tabulation of number of pictures tagged in each modality for participants in the Both condition

subject numb.	Audio	Text	Both	Total numb. pict.
14	0	**151**	19	170
15	14	1	**41**	56
18	**99**	3	0	102
20	7	**31**	4	42
5	10	**43**	12	65

tagging and retrieval modality. *Before* the controlled experiment took place, 4 participants reported considering that text tags were more effective to retrieve images than audio tags, while the remaining two participants did not have a specific opinion. However, *after* the experiment *all* participants in the BOTH group felt that tagging with text was more effective than tagging with voice (Likert scale: 1=completely disagree – 5=completely agree. Voice: 3.33 [0.81], Text: 4.34 [0.81]). These results contradict H1-bis, which was predicting the users' preference for speech annotations.

In addition, participants in the TEXT condition thought that tagging with text was effective (4.33 [0.52]), but uncomfortable while on the move (2.00 [1.15]). Participants in the VOICE condition were somewhat indifferent about the effectiveness of tagging with voice (3.14 [1.06]) and felt uncomfortable tagging with voice in public (2.75 [1.38]).

Concerning **H2**, it was found that participants in the VOICE condition took a shorter amount of time to produce their tags compared to participants in the TEXT condition (Voice: 2.58[.84] *vs.* Text: 7.04[.96] *vs.* Both: 6.55 [4.44] sec., F[2, 16] = 5.67, p<.05). The post-hoc test revealed that the difference between the TEXT and the BOTH condition was not significant (p>0.9, ns). In order to study in more detail the relation between the production time of the tag and the length of the tag itself, the length of the audio tag and the number of characters of the text tag were converted into z-scores. This was necessary to make these two measures comparable. Then a Pearson's correlation test was conducted between the time required to create a tag and the length z-score previously computed. The test revealed a medium correlation between the two measures (N=1,9345, r=.355, p<.001). The plot in Figure 4 demonstrates this relation. The slope of the linear fit of the text tags is steeper than the slope of the linear fit for the audio tags. In other words, participants tagging with text took *longer* to produce their tags than participants tagging with voice. Thus, these results verify H2, which was predicting an increasing advantage when entering tags with voice in relation to the length of the tag.

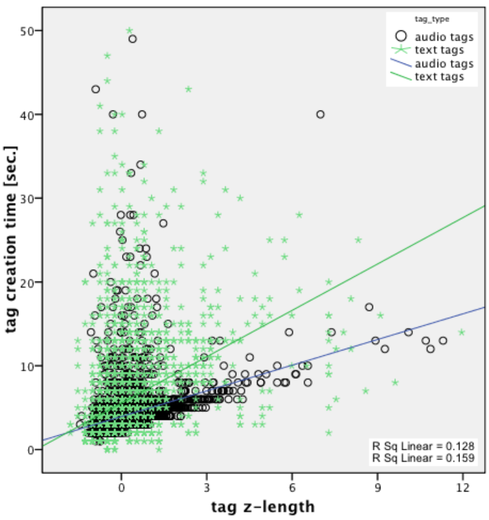

Figure 4: Scatterplot of the tag creation time vs. tag z-length

In summary, these results demonstrate that *in the case of long tags, participants that tagged with voice spent less time tagging than those that tagged with text.* Finally, participants who could tag their pictures with text and audio, did not show a preference for either modality.

5.2 Controlled Experiment

Participants took a different number of pictures during the field experiment. In order to avoid the potential impact of this variable on task completion time, all participants were given the exact same number of pictures during the retrieval experiment, namely 150 pictures[5].

The controlled experiment was designed to test the third hypothesis (**H3**). It was found that the experimental condition had an effect on the time required by the participants to solve task T1. While solving the first task, participants in the VOICE condition took longer than the time required by participants in the TEXT condition (Voice: 29.71 [8.56] *vs.* Text: 18.03 [6.50] *vs.* Both: 17.94 [4.45] sec., F[2, 16] = 5.69, p<.05). The post-hoc test revealed that the difference between the TEXT and the BOTH condition was not significant (p>0.9, ns). No significant effects of the condition were found on the task completion times for T2, T3 and T4. These results, summarized in Table 2, confirm H3,

[5] Note that all participants took at least 150 pictures during the field study.

which was predicting equal time-to-completion for subjects in the different experimental conditions.

Table 2: Time-To-Completion for T1–T4, Mean [std], sec., (*p<.05). Descriptive statistics with different subscripts (*i.e.* a, b) in the same row diverge significantly for p<.05

	Voice	Text	Both	P
T1	29.71[8.56]$_a$	18.03[6.50]$_b$	17.94[4.45]$_b$	*
T2	33.88[13.20]	40.20[34.57]	36.57[18.02]	ns
T3	30.50[22.18]	26.66[18.31]	24.82[10.48]	ns
T4	17.11[6.70]	20.56[7.07]	18.05[7.96]	ns

Additionally, we found the experimental condition to have an effect on the number of false positive answers and retrieval errors incurred by the participants for T1, T2, and T3. However, there was no significant difference in the number of queries executed by participants to solve the task. These results are summarized in Table 3. Participants in the Voice condition provided more false positive answers than participants in the Text or Both conditions (Voice: 3.17 [1.17] *vs.* Text: .83 [.75] *vs.* Both: 1.80 [1.30] answers, F[2, 16] = 7.00, p<.05). The post-hoc test revealed that the differences among the Text and Both condition were not significant when taken two by two (p>.05, ns). Similarly, participants in the Voice condition made more retrieval errors than participants in the Text or Both conditions (Voice: 3.33 [.82] *vs.* Text: 1.00 [1.27] *vs.* Both: 2.20 [1.10] answers, F[2, 16] = 7.09, p<.05). The post-hoc test revealed that the differences among the Text and Both condition were not significant when taken two by two (p>.05, ns).

Table 3: Number of: False positives, queries and retrieval mistakes for T1–T3, Mean [std], (*p<.05). Descriptive statistics with different subscripts (*i.e.* a, b) in the same row diverge significantly for p<.05

	Voice	Text	Both	P
False P.	3.17[1.17]$_a$	.83[.75]$_b$	1.80[1.30]$_b$	*
Queries	12.77[3.64]	10.94[4.24]	13.93[2.96]	ns
Retr. Err.	3.33[.82]$_a$	1.00[1.27]$_b$	2.20[1.10]$_b$	*

In order to better understand the differences across the experimental conditions, we computed a cross-tabulation of the absolute frequencies of false positives and retrieval mistakes by the experimental condition (see Table 4). The results suggest that participants in the Voice condition had more difficulties in remembering the tag associated with a certain picture, which was requested by T1 and T2. Additionally, these participants had more difficulty in remembering the picture associated with a certain tag, which was requested by T2 and T3.

In summary, these results show that subjects who tagged their pictures with text tags finished the tasks in the same amount of time as subjects who tagged their pictures with audio, with the exception of the first task, which was solved faster by participants who tagged their pictures with text. In addition, participants in the Text condition made fewer mistakes, as reflected by fewer false positive answers and retrieval errors. Finally, participants who had at their disposal both Text and Voice tags behaved similarly to participants who had only Text tags.

Table 4: Frequencies of false positives and retrieval mistakes in each task by experimental condition

	Voice				Text				Both			
	T1	T2	T3	T4	T1	T2	T3	T4	T1	T2	T3	T4
False P.	0	**8**	**11**	23	0	**3**	**2**	23	0	**6**	**3**	18
Retr. Err.	**6**	**10**	4	0	**1**	**4**	1	0	**3**	**6**	2	0

6. DISCUSSION

Speech has been proposed in previous work as a suitable modality in the interaction with mobile devices. However, the results of the study presented in this paper caution that the advantage of audio as an input modality for tagging pictures on mobile phones is not a given. The objective measures of the tagging experiment show that those who could choose between speech and text as input modalities did not use preferably the former as predicted by **H1** (suggested by previous work [8, 9]).

The analysis of the subjective responses showed that the majority of participants found textual tags to be more effective than voice tags, both for tagging and retrieval, thus contradicting **H1-bis**, which was predicting a subjective preference for voice. This result is not consistent with the results of Perakakis and Potamianos [18], possibly because part of our study was conducted on the field where participants could experience speech input in real-life situations. When explaining *why* they preferred text over voice, participants indicated two reasons: (a) *retrieval precision* and (b) *privacy*: (a) the audio tags were recorded with random background noises, which lead to increased retrieval errors; and (b) participants felt sometimes uncomfortable speaking to their phones in public places.[6]

> (1, Both) It is easier for me to browse using text that using voice. When labeling with voice, I am not sure if the label has been recorded properly or not, as it can also contain background noise and other factors that affect the recording. Conversely, a word is always written in the same way.
> (2, Voice) I felt embarrassed about tagging with voice, because there were people listening to me. It was a little slow and bothersome to tag each picture and insert several tags.

In the study, it was found that long tags were entered via voice faster than when using a mobile phone keypad, as predicted by **H2** (suggested by Hauptmann and Rudnicky [9]). It was found that participants tagging with text spent more time editing their tags. In addition, the longer the tag, the longer it took participants to type it when compared to speech. One can expect this relation to be even more intense for users using a 12-keys pad instead of the screen keyboard our subjects used. Note how the green stars in figure 4 –corresponding to the text modality– are more sparsely distributed than the black circles –that belong to the audio modality. Possibly, this is due to the fact that users tend to immediately address input mistakes (*e.g.*, typos) while typing. Therefore, pausing in the midst of textual input is not

[6] The following quotes from the questionnaires were translated from Spanish to English. Each excerpt is marked with a progressive number and the experimental condition of the subject.

uncommon. On the contrary, these real-time corrections are very rare in the case of voice input.

Speech offered a clear advantage over text in terms of the time that it took to compose a tag. However, the same advantage could not be found during the retrieval phase, as predicted by **H3** (suggested by the work of Rudnicky [20] and Mills et al. [15]). Participants who tagged with voice did not solve the controlled tasks faster than participants who tagged with text. Furthermore, it was found the opposite relation for T1, where those who tagged with text or with text and voice solved the task quicker than those who tagged with voice. Note that T1 required remembering the tags that were used when indexing the picture. The experimental results suggest that participants in the Voice condition had a harder time remembering the association between pictures and tags. This observation is supported by the higher number of *false positives* and *retrieval errors* for all the tasks in the Voice condition when compared to the other two conditions. Indeed, this might be caused by the shorter amount of time that participants spent introducing the tag and by the lack of visual feedback, which might have had an impact on the memorization of the tag itself [4].

> (3, Text) Upon starting the experiment I thought it would had been better adding tags via voice (I imagined it to be more practical and/or convenient). However, I think that text tags are better for searching, even though they might be sometimes a bit boring and bothersome to enter. They (*i.e.* text tags) force you to think more about what tag to assign and the way to find the most adequate word. They help to memorize more easily the assigned tag to a given picture.

With respect to designing mobile applications, speech as input modality offers different and complementary advantages over text that are well recognized in the literature: (a) It does not require both hands, which is probably very useful while on the move; (b) speech is not influenced by lighting conditions and glare; and (c) speech does not require precise fine motor skills as it is the case when typing on a small keypad. However, the findings of this study help reconsider the importance of the textual modality for entering tags in relation to the retrieval precision and privacy of the user.

All together, these results provide an answer to our first research question: *The production of picture tags on mobile devices is helped by voice tags, which reduce the editing time and the effort required from the user. However, the major limitation of this modality is that of exposing the user's privacy. Conversely, retrieval is helped by textual tags, which are easier to memorize and more precise to enter and match with existing tags, providing consistent results.* In the following, we address our second research question.

6.1 Implications for Design

The results of the experiment suggest the following guidelines for the design of mobile multimodal tagging applications:

1. **Allow multiple modalities**. In the case of a single modality, text should be preferred over voice unless the tagging context does not violate the user's privacy. Additionally, great advantages can be achieved by integrating both modalities (*e.g.*, If I am in a crowded place and I have time, I might use text; otherwise, I will use voice). Moreover, applications and users would benefit even more by enabling *interoperability between the modalities.* Given the state-of-the-art in speech processing, we believe that this might be achieved by post-process speech-to-text and text-to-speech conversion. The benefits of the interoperability of modalities include:

a) *Once converted, tags might allow the user to execute cross-modality queries.* In fact, many users in the Both condition, reported not remembering the modality that they recorded a certain tag with. Conversion of the tags to the same modality would allow users to execute queries in voice or text indistinctively.

b) *The tags could be easily reused across different services and reused by other users*, which would minimize the effort and maximize the incentives –as Kustaniwitz and Shneiderman [13] proposed. Note that this is not the case with speech tags that are not converted to text, as they are closely related to their producer and hence less likely to be reused by other users. Furthermore, annotations from the digital camera phone could conveniently/directly map to those on the social networking sites.

2. **Enable audio inspection**. Audio recordings cannot be easily inspected. Therefore, the user might not be aware of "audio typos" such as background noise or half recorded sentences. Hence, it is extremely important to provide feedback about the quality of the audio recording. This will probably increase the confidence of the user in the system and allow better re-execution of the same voice tag during retrieval.

3. **Modality Synesthesia**. As text tags seem to be easier to remember –partially because of their visual feedback, the user might benefit from an algorithm that would dynamically associate his/her speech tag with a visual representation that would act as redundant feedback to help memorization. This visual feedback would be displayed right after the speech is captured.

4. **Enable tagging deferral and improvements of tagging**. There is clearly an advantage in tagging when the image is captured. However, tagging takes time and users might not always be willing to spend the time to create the tags.

> (4, Text) Personally, I would have preferred tagging the images once they had been taken and not right after shooting. Sometimes it is not convenient, or you don't have much time to take many pictures at once.

Therefore, users should be given the option to tag the pictures at a later time. Additionally, semi-automatic tagging mechanisms might help to reduce the overhead of tagging sequences of pictures taken in the same place and at about the same time (see for instance the techniques proposed by ZoneTag [1] or the other proposal based on the use of retrieval queries [25]). In the scenario presented at the beginning of this paper, it might be enough to support tagging and retrieval of entry points to the sequence of pictures from where users might start a slideshow. More research would need to be conducted in order to demonstrate the ability of the user to remember these *entry points* (*i.e.*, pictures with associated tags).

7. CONCLUSIONS

In this paper, we have presented a comparison of textual and speech annotation modalities for pictures captured with camera phones. The findings suggest that each tagging modality has advantages in specific contexts: voice tags are preferred while on the run, and text tags when users are in public and potentially crowded places. Additionally, voice tags do not seem to offer specific advantages when compared to textual tags for retrieving pictures. Furthermore, the results of the first task in this experiment suggest that textual tags are more effective for retrieval as they are easier to remember than speech tags.

When considering speech-to-text techniques instead of the approach used in MAMI, even higher preference for textual tags over voice tags could be expected due to transcription problems [5, 15]. However, considering optimal improvements on such algorithms (*e.g.* faster processing, semantic associations, *etc.*), user studies like this could yield different results, given the advantages of speech for mobile devices [10, 18]. In fact, the experiments presented in this paper have revealed that voice tags are entered faster than text tags, as subjects tend to spend more time editing textual tags (in agreement with Cox and colleagues [7]). Some participants also mentioned their preference for adding tags with voice and retrieving the pictures with text. In this sense, a multimodal approach for mobile tagging applications seems very relevant for a better user experience (supported by Paek and colleagues [17]). It is important to highlight that in the presented experiment, the combination of voice and text inputs was not designed on the basis of interaction patterns and therefore it does not tell us how a better designed integration of the two modalities could work.

Additional guidelines in the design of multimodal mobile image tagging and retrieval applications include: defer data processing to idle times, postpone tag annotation to a later time than that of the image capture and allow audio inspection in order to increase the user's confidence on their speech annotations.

To conclude, the main result of this work is somewhat intuitive yet important: each input modality has advantages and specificities that influence uses and performance. Great possibilities lie in the wise integration of modalities in mobile phones.

Acknowledgments We would like to thank all the participants of the study for donating their holiday time to science and sharing their mementos of their life and their thoughts with us. Finally, we thank Enrico Costanza, Fernando Flores-Mangas, and Karen Church for reviewing an early draft of this paper.

8. REFERENCES

[1] M. Ames and M. Naaman. Why we tag: motivations for annotation in mobile and online media. In *Proc. CHI 2007*, pages 971–980. ACM Press (2007).

[2] X. Anguera, N. Oliver, and M. Cherubini. Multimodal and mobile personal image retrieval: A user study. In K. L. Chan, editor, *Proc. MobIR 2008*, pages 17–23.

[3] X. Anguera, J. Xu, and N. Oliver. Multimodal photo annotation and retrieval on a mobile phone. In *Proc. MIR 2008*, pages 188–194. ACM Press (2008).

[4] A. D. Baddeley. *Human Memory: Theory and Practice.* Psychology Press, London, UK, 1997.

[5] M. G. Brown, J. T. Foote, G. J. F. Jones, K. S. Jones, and S. J. Young. Open-vocabulary speech indexing for voice and video mail retrieval. In *Proc. Multimedia'96*, pages 307–316. ACM Press (1996).

[6] J. Chen, T. Tan, and P. Mulhem. A method for photograph indexing using speech annotation. In *Proc. PCM 2001*, pages 867–872. Springer-Verlag Press (2001).

[7] A. Cox, P. Cairns, A. Walton, and S. Lee. Tlk or txt? using voice input for sms composition. *Pers. and Ubiq. Computing*, 12(8):567–588, 11 2008.

[8] S. Hah and V. Ahlstrom. Comparison of speech with keyboard and mouse as the text entry method. In *Proc. of the Human Factors and Ergonomics Society*, pages 619–622, 2005.

[9] A. G. Hauptmann and A. I. Rudnicky. A comparison of speech and typed input. In *Proc. HLT 1990*, pages 219–224. ACL Press (1990).

[10] T. J. Hazen, B. Sherry, and M. Adler. Speech-based annotation and retrieval of digital photographs. In *Proc. INTERSPEECH 2007*, pages 2165–2168.

[11] T. Kindberg, M. Spasojevic, R. Fleck, and A. Sellen. The ubiquitous camera: An in-depth study of camera phone use. *IEEE Pervasive Computing*, 4(2):42–50, 2005.

[12] A. Kuchinsky, C. Pering, M. L. Creech, D. Freeze, B. Serra, and J. Gwizdka. FotoFile: a consumer multimedia organization and retrieval system. In *Proc. CHI 1999*, pages 496–503. ACM Press (1999).

[13] J. Kustaniwitz and B. Shneiderman. Motivating annotation for personal digital photo libraries: Lowering barriers while raising incentives. Technical Report HCIL-2004-18, University of Mariland, 2005.

[14] K. M. Lee and J. Lai. Speech versus touch: A comparative study of the use of speech and dtmf keypad for navigation. *Int. Journal of Human-Computer Int.*, 19(3):343 – 360, 2006.

[15] T. J. Mills, D. Pye, D. Sinclair, and K. R. Wood. Managing photos with AT&T Shoebox (demo session). In *Proc. SIGIR 2000*, page 390. ACM Press (2000).

[16] H. Mitchard and J. Winkles. Experimental comparisons of data entry by automated speech recognition, keyboard, and mouse. *Human Factors*, 44(2):198–209, Summer 2002.

[17] T. Paek, B. Thiesson, Y.-C. Ju, and B. Lee. Search vox: leveraging multimodal refinement and partial knowledge for mobile voice search. In *UIST '08*, pages 141–150, New York, NY, USA, 2008. ACM.

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

[19] K. Rodden and K. R. Wood. How do people manage their digital photographs? In *Proc. CHI 2003.*

[20] A. I. Rudnicky. Mode preference in a simple data-retrieval task. In *Proc. HLT 1993*, pages 364–369. ACL Press (1993).

[21] Herrarte. E. Wilhelm. A. Sarvas, R. and M. Davis. Metadata creation system for mobile images. In *Proc. MobiSys'04*, (2004).

[22] R. K. Srihari. Multimedia indexing and retrieval of voice-annotated consumer photos. In *Proc. SIGIR 1999*, pages 1–16. ACM.

[23] A. Stent and A. Loui. Using event segmentation to improve indexing of consumer photographs. In *Proc. SIGIR 2001*, pages 59–65. ACM Press (2001).

[24] T. Tan, J. Chen, P. Mulhem, and M. Kankanhalli. Smartalbum: a multi-modal photo annotation system. In *Proc. Multimedia 2002*, pages 87–88. ACM Press (2002).

[25] L. Wenyin, S. Dumais, Y. Sun, H. Zhang, M. Czerwinski, and B. Field. Semi-automatic image annotation. In *Proc. INTERACT 2001*, pages 326–333.