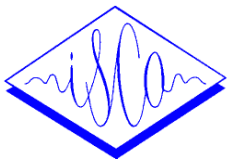


6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

MIPAD: A NEXT GENERATION PDA PROTOTYPE

X. Huang, A. Acero, C. Chelba, L. Deng, D. Duchene, J. Goodman, H. Hon, D. Jacoby, L. Jiang, R. Loynd, M. Mahajan, P. Mau, S. Meredith, S. Mughal, S. Neto, M. Plumpe, K. Wang, Y. Wang

Speech Technology Group
Microsoft Research
Redmond, Washington 98052, USA
http://research.microsoft.com/stg

ABSTRACT

MiPad is one of the application prototypes in a project codenamed Dr Who. As a wireless Personal Digital Assistant (PDA), MiPad fully integrates continuous speech recognition (CSR) and spoken language understanding (SLU) to enable users to accomplish many common tasks using a multimodal interface and wireless technologies. It tries to solve the problem of pecking with tiny styluses or typing on minuscule keyboards in today's PDAs or smart phones. It also avoids the problem of being a cellular telephone that depends on speech-only interaction. MiPad incorporates a built-in microphone that activates whenever a field is selected. As a user taps the screen or uses a built-in roller to navigate, the tapping action narrows the number of possible instructions for spoken language processing. MiPad currently runs on a Windows CE Pocket PC with a Windows 2000 Server where speech recognition is performed. The Dr Who CSR engine has a 64k word vocabulary with a unified context-free grammar and n-gram language model. The Dr Who SLU engine is based on a robust chart parser and a plan-based dialog manager. This paper discusses MiPad's design, implementation work in progress, and preliminary user study in comparison to the existing pen-based PDA interface.

1. INTRODUCTION

There are three broad classes of applications that our Dr Who project is trying to address:

- Office: This is the widely used desktop application such as Microsoft Windows and Office.
- Home: TV and kitchen are the center for home application. Since home appliances and TV don't have a keyboard or mouse, the traditional GUI application can't be directly extended for this category.
- Mobile: Cell phone and car are two most important mobile scenarios. Because the physical size and hands-busy and eyes-busy constraints, the traditional GUI application interaction model requires a significant modification.

Spoken language has the potential to provide a consistent and unified interaction model across these three classes, albeit for these different application scenarios, you still need to apply different user interface (UI) design principles. MiPad is one of Dr Who's applications that addresses the mobile interaction scenario. It is a wireless PDA that enables users to accomplish many common tasks using a multimodal spoken language interface (speech + pen + display) and wireless-data technologies. This paper discusses MiPad's design, implementation work in progress, and preliminary user study in comparison to the existing pen-based PDA interface. Several functions of MiPad are still in the designing stage, including its hardware design. One of its hardware design concepts is illustrated in Figure 1.

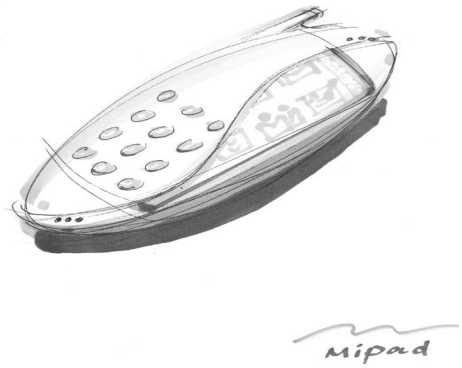


Figure 1 One of MiPad's industrial design concepts

MiPad tries to solve the problem of pecking with tiny styluses or typing on minuscule keyboards in today's PDAs. It also avoids the problem of being a cellular telephone that depends on speech-only interaction. It has a built-in microphone that activates whenever a visual field is selected. MiPad is designed to support a variety of tasks such as E-mail, voice-mail, Web browsing, cellular phone. This collection of functions unifies the various devices that people carry around today into a single, comprehensive communication tool. While the entire functionality of MiPad can be accessed by pen alone, it can also be accessed by speech and pen combined. The user can dictate to a field by holding the pen down in it. The pen simultaneously acts to focus where the recognized text goes, and acts as a push-to-talk control. As a user taps the screen or uses a built-in roller to navigate, the tapping action narrows the number of possible instructions for spoken language processing.

Currently, we only implemented MiPad's Personal Information Management (PIM) functions: email, calendar, and contact list. MiPad's hardware prototype is based on Compaq's iPaq. It is configured with a client-server architecture as shown in Figure 2. The client is based on Microsoft Windows CE that contains only signal processing and UI logic modules. The wireless local area network (LAN), which is currently used to simulate wireless 3G, connects the client to a Windows 2000 Server where CSR and SLU are performed. The bandwidth requirement between the signal processing module and CSR engine is about 2.5-4.8kbps. At 2.5-4.8 kbps, we observed less than 5% relative error increase for the CSR engine due to the parameter compression. MiPad

applications communicate via our dialog manager to both the CSR and SLU engines for coordinated context-sensitive *Tap and Talk* interaction.

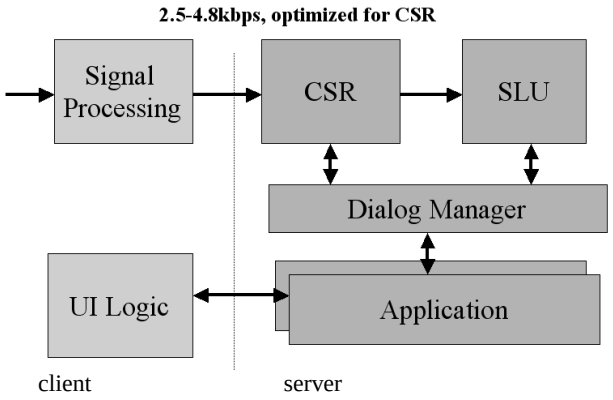


Figure 2 MiPad's client-server architecture. The client is based on a Windows CE iPAQ, and the server is based on a Windows 2000 server. The client-server communication is currently based on the wireless LAN.

2. MIPAD UI DESIGN

2.1 Tap and Talk interface

Since MiPad is a small handheld device, the present pen-based methods for getting text into a PDA (Graffiti, Jot, soft keyboard) are barriers to broad market acceptance. As an input modality, speech is generally not as precise as mouse or pen to perform position-related operations. Speech interaction can be adversely affected by the ambient noise. When privacy is of concern, speech is also disadvantageous since others can overhear the conversation. Despite these disadvantages, speech communication is not only natural but also provides a powerful complementary modality to enhance the pen-based interface. Because of these unique features, we need to leverage the strengths and overcome the technology limitations that are associated with the speech modality. As shown in Table 1, pen and speech can be complementary and they can be used very effectively for handheld devices. You can tap to activate microphone and select appropriate context for speech recognition. The advantage of pen is typically the weakness of speech and vice versa. This implied that user interface performance and acceptance could increase by combining both. Thus, visible, limited, and simple actions can be enhanced by nonvisible, unlimited and complex actions.

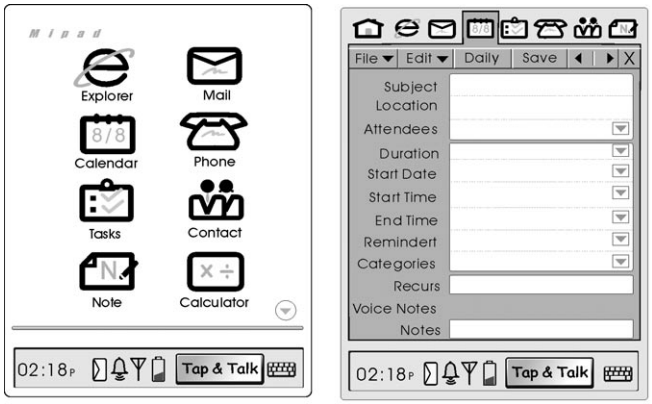
Table 1 Complementary strengths of pen and speech as input modalities

Pen	Speech
Direct manipulation	Hands/eyes free manipulation
Simple actions	Complex actions
Visual feedback	No Visual feedback
No reference ambiguity	Reference ambiguity

People tend to like to use speech to enter data and pen for corrections and pointing. As illustrated in Table 2, MiPad's *Tap and Talk* interface offers a number of benefits. MiPad has a *Tap & Talk* field that is always present on the screen as illustrated in MiPad's start page in Figure 3 (a) (the bottom gray window is always on the screen).

Table 2 Benefits to have speech and pen for MiPad

Action	Benefit
Ed uses MiPad to read an e-mail, which reminds him to schedule a meeting. Ed taps to activate microphone and says *Meet with Peter on Friday.*	Using speech, information can be accessed directly, even if not visible. Tap and talk also provides increased reliability for speech detection.
The screen shows a new appointment to meet with Peter at 10:00 on Friday for an hour.	An action and multiple parameters can be specified in only a few words.
Ed taps Time field and says *Noon to one thirty*	Field values can be easily changed using field-specific language and semantic models
Ed taps Subject field dictates and corrects a couple of sentences explaining the purpose of the meeting.	Bulk text can be entered easily and faster.



(a) (b)

Figure 3 Concept design for (a) MiPad's first card and (b) MiPad's calendar card

The user can give spontaneous commands by tapping the *Tap & Talk* field and talking to it. The system recognizes and parses the command, such as showing a new appointment form as illustrated in. The appointment form shown on MiPad's display is similar to the underlying semantic objects. The user can have conversation by *tapping and talking* to any sub-field as well. By tapping to the attendees field in the calendar card shown in Figure 3 (b), for example, the semantic information related to potential attendees is used to constrain both CSR and SLU, leading to a significantly reduced error rate and dramatically improved throughput. This is because the perplexity is much smaller for each subfield-dependent language and semantic model.

2.2 Fuzzy soft keyboard

Since we have a language model for speech recognition, we can use the same knowledge source to reduce the error rate of the soft keyboard when it is used instead of speech recognition. We model the position of the stylus tap as a continuous variable, allowing the user to tap either in the intended key, or perhaps nearby in an adjacent key. By combining this position model with a language

model, error rates can be reduced. In our preliminary user study, the average user made half as many errors on the fuzzy soft keyboard, and almost all users preferred the fuzzy soft keyboard.

3. SPOKEN LANGUAGE PROCESSING

3.1 Acoustic modeling

Since MiPad is a personal device, we can use speaker-adaptive acoustic modeling for improved speech recognition. The Dr Who CSR engine is an improved version of Microsoft's Whisper speech recognition system [2]. Both MLLR and MAP adaptation are used to adapt the speaker-independent acoustic model for each individual speaker. There are 6000 senones with 20-mixture continuous Gaussian densities. The context-sensitive language model is used for relevant semantic objects driven by the user's pen tapping action, as described in the MiPad's *Tap and Talk* interface design.

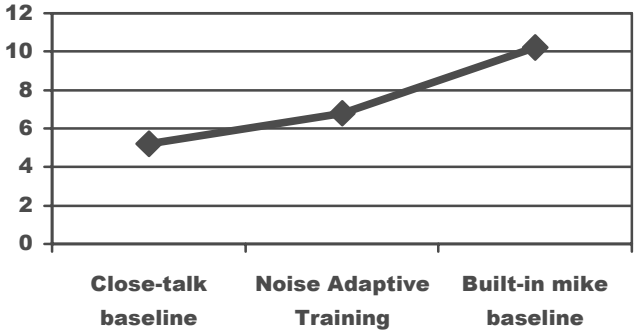


Figure 4 Word recognition error rates of close-talk microphone and built-in microphone with or without noise adaptive training.

In the typical MiPad usage scenario, the user may use the built-in MiPad microphone that is very sensitive to environment noise. When we evaluated the built-in microphone of Compaq's iPaq device, the word recognition error rate increased by a factor of two in comparison to a close-talk microphone (in the normal office environment), which shows that the close-talk microphone is necessary despite its inconvenience. Since this error increase is mainly due to the additive environment noise, the Dr Who CSR engine used our noise adaptive training to improve the performance of the built-in microphone [1]. As shown in Figure 4, performance with the built-in microphone can be dramatically improved with the noise adaptive training technique. Here the test data used for evaluation are based on the WSJ dictation task.

3.2 Language modeling

The Dr Who CSR engine uses the unified language model [5] that takes advantage of both rule-based and data-driven approaches.

Consider two training sentences: *Meeting at three with Zhou Li.* vs. *Meeting at four PM with Derek.* If we use a word trigram, we will estimate *P(Zhou|three with)* and *P(Derek|PM with)*. There is no easy way to capture needed long-span semantic information in the training data. The unified model uses a set of CFGs that can capture the semantic structure of the domain. For the example listed here, we may have a CFG for {name} and {time} respectively, which can be derived from the natural language parser in the training data. The training sentences now look like: *Meeting {at three:TIME} with {Zhou Li:NAME}.* and *Meeting {at four PM:TIME} with {Derek: NAME}.* With parsed training data, we can estimate our n-gram probabilities as usual. We have probabilities such as P({name}|{time} with) instead of *P(Zhou|three with)*, which is more meaningful and accurate. Inside each CFG, we can also derive P("Zhou Li"|{name}) and P("four PM"|{time}) from the existing n-gram (n-gram probability inheritance) so that they are normalized [5]. The unified approach can be regarded as a standard n-gram in which the vocabulary consists of words and structured classes. The structured class can be very simple such as {date}, {time}, and {name} or can be very complicated such as a CFG that contains deep structured information. The key advantage of the unified language model is that we can author limited CFGs for each new domain and embed them into the domain independent n-gram model.

Most decoders can only support either CFGs or word n-grams. We have modified our decoder so that we can embed CFGs in the n-gram search framework to take advantage of our unified language model. As shown in Table 3, the unified language model significantly improves cross-domain portability.

The test data shown here are based on MiPad's PIM *conversational speech*. The domain-independent trigram language model is based on Microsoft Dictation trigram models used in Microsoft Speech SDK 4.0. From the table, we can see that it is important to use the unified model in the early stage, which outperformed results based on lattice re-scoring.

Table 3 Cross-domain speaker-independent speech recognition performance with the unified language model and its corresponding decoder

Systems	Perplexity	Word Error	~Time
Domain-independent Trigram	593	35.6%	1.0
Unified decoder with the unified LM	141	22.5%	0.77
N-best re-scoring with the unified LM	-	24.2%	-

3.3 Spoken language understanding

The Dr Who SLU engine is based on a robust chart parser [4] and a plan-based dialog manager [3]. Each semantic class is associated with an action that the application takes. Each semantic class has slots that require a context-free grammar. These semantic objects are mapped to the graphic card the user can see directly on MiPad's display. When appropriate semantic objects are decided, the dialog manager decides the flow of these semantic objects, which includes both inter and intra-frame control and error repair strategy.

One of the critical tasks is semantic grammar authoring. It is necessary to collect a large amount of real data to author the semantic grammar to reach a decent coverage. Even for the PIM subtasks, we found that the Dr Who SLU engine's slot parsing error rate in the general *Tap and Talk* field is above 40%. This result is obtained from filed-independent sentences (i.e. from the

general-purpose *Tap & Talk* field). About half of these errors are due to the free-form text that are related to email or meeting subjects.

After collecting additional MiPad data, we are able to reduce the SLU error by more than 25%, which is still insufficient to be useful. Fortunately, with our imposed context constraints in the *Tap and Talk* interface, where field-specific language and semantic models can be used, we can overcome most of today's SLU technology limitations.

4. USER STUDY RESULTS

It is our ultimate goal to make sure that Dr Who technologies add value to our customers. It is necessary to have a rigorous evaluation to measure the usability of the MiPad prototype. Our major concerns are *Is the task completion time much better?* and *Is it easier to get the job done?*

For our preliminary user study, we set out to assess the performance of the current version of MiPad (with PIM features only) in terms of task-completion time (for both CSR and SLU), text throughput (CSR only), and user satisfaction. The focal question of this study is whether the *Tap and Talk* user interface can provide added value to the existing PDA user interface.

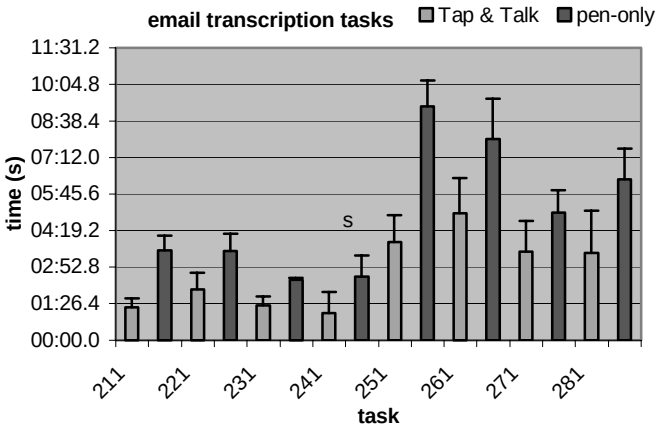


Figure 5 Task completion time of email transcription between the pen-only interface and *Tap and Talk* interface. The standard deviation is also shown above the bar of each performed task.

Is the task completion time much better? 20 computer-savvy users tested the partially implemented MiPad prototype. These people had no experience with PDAs or speech-recognition software. The tasks we evaluated include creating a new email, checking calendar, and creating a new appointment. Task order was randomized. We alternated tasks for different user groups using either pen-only or *Tap and Talk* interfaces. The text throughput is calculated during e-mail paragraph transcription tasks. Compared to using the pen-only user interface, we observed that the *Tap and Talk* interface is about 50% faster transcribing email documents[1]. For the overall command and control operations such as scheduling appointments, the *Tap and Talk* interface is about 33% faster than the existing pen-only interface[2]. Error correction for the *Tap and Talk* interface remains as one of the most unsatisfactory features. In our user study, calendar access time using the *Tap and Talk* methods is about the same as pen-only methods, which suggests that simple actions are very suitable for pen-based interaction.

Is it easier to get the job done? Most users we tested stated that they preferred using the *Tap and Talk* interface. The preferences are consistent with the task completion times. Indeed, most users comments concerning preference were based on ease of use and time to complete the task, as demonstrated in Figure 5.

5. SUMMARY

MiPad is a work in progress for us to develop a consistent Dr Who interaction model and Dr Who engine technologies for three broad classes of applications. A number of discussed features are yet to be fully implemented and tested. Our currently tested features include PIM functions only. Despite our incomplete implementation, we observed that speech and pen have the potential to significantly improve user experience in our preliminary user study. Thanks to the multimodal interaction, MiPad also offers a far more compelling user experience than standard telephony interaction.

The success of MiPad depends on spoken language technology and always-on wireless connection. With upcoming 3G wireless deployments in sight[3], the critical challenge for MiPad remains the accuracy and efficiency of our spoken language systems since likely MiPad may be used in the noisy environment without using a close-talk microphone, and the server also needs to support a large number of MiPad clients.

ACKNOLWEDEGEMENT

We thank E. Chang, M. Czerwinski, J. Breese, D. Ling, X. Lu, K. Steury, and D. Venolia, for their help in Dr Who's R&D.

REFERENCES

[1] Deng, L., *et al. Large-Vocabulary Speech Recognition Under Adverse Acoustic Environments*. in *ICSLP*. 2000. Beijing, China.

[2] Huang, X., *et al., From Sphinx II to Whisper: Making Speech Recognition Usable*, in *Automatic Speech and Speaker Recognition*, C.H. Lee, F.K. Soong, and K.K. Paliwal, Editors. 1996, Klewer Academic Publishers: Norwell, MA. p. 481-508.

[3] Wang, K. *A Plan-Based Dialog System With Probabilistic Inferences*. in *ICSLP*. 2000. Beijing, China.

[4] Wang, Y. *A Robust Parser For Spoken Language Understanding*. in *Eurospeech*. 1999. Hungary.

[5] Wang, Y., M. Mahajan, and X. Huang. *A Unified Context-Free Grammar And N-Gram Model For Spoken Language Processing*. in *International Conference on Acoustic, Signal and Speech Processing*. 2000. Istanbul, Turkey.

[1] The corresponding speaker-adaptive speech recognition error rate for the email transcription tasks is about 14%, which is based on using a close-talk microphone and a speaker-adaptive acoustic model trained from about 20 minutes speech.

[2] The speaker-dependent SLU error rate for different cards (not slots) is about 4%.

[3] http://www.wirelessweek.com/issues/3G