MIPAD: A MULTIMODAL INTERACTION PROTOTYPE

X. Huang, A. Acero, C. Chelba, L. Deng, J. Droppo, 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. Steury, G. Venolia, K. Wang, Y.Wang

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

ABSTRACT

Dr. Who is a Microsoft's research project aiming at creating a speech-centric multimodal interaction framework, which serves as the foundation for the .NET natural user interface. MiPad is the application prototype that demonstrates compelling user advantages for wireless Personal Digital Assistant (PDA) devices, 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. Unlike a cellular phone, MiPad avoids speech-only interaction. It 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 understanding. MiPad currently runs on a Windows CE Pocket PC with a Windows 2000 machine where speech recognition is performed. The Dr Who CSR engine uses a unified CFG 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

While graphic user interface (GUI) significantly improve man-machine interface by using intuitive real-world metaphors, it is still far away from ultimate goal where users can interact with any system without any training. Particularly, GUI relies heavily on a sizeable screen, keyboard and pointing device; whereas the sizeable screen, keyboard or pointing device is not available. There are two broad classes of applications that Dr Who project is trying to address:

- Home: TV and kitchen are the center for home application. Since home appliances and TV don't have a keyboard or mouse, the GUI interaction could be awkward to use.
- Mobile: Cell phone and car are two most important mobile scenarios. Because the physical size and hands-busy and eyes-busy constraints, the GUI interface face even bigger challenge.

While spoken language has the potential to provide a natural interaction model, the ambiguity of spoken language and the memory burden of using speech as output modality on the user prevent it becoming the choice of mainstream interface.

Multimodality that is a normal interaction model for human-human communication, is thought to be capable of dramatically enhancing the usability of speech because GUI and speech have complementary strengths. Dr. Who is Microsoft's attempt to develop a speech-centric multimodal interface framework and related enabling technologies. MiPad is the first 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 technologies. This paper discusses MiPad's design, implementation work, 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.

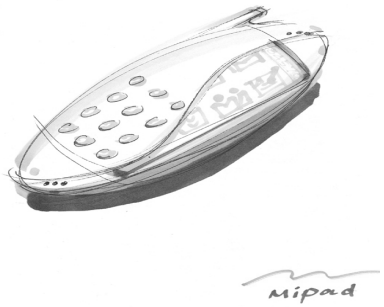


Figure 1 One of MiPad's industrial design concepts

MiPad tries to solve the problem of pecking with tiny styluses or typing on minuscule keyboards in today's PDAs. Unlike a cellular phone, MiPad avoid 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, calendar, and web browsing. While the entire functionality of MiPad can be accessed by pen alone, it is preferred to 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, contact list, and memos. 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 a Microsoft Windows CE application that

contains only front-end processing and UI logic modules, and a robust communications layer that allows the system to recover gracefully from the connection failures of an unreliable cellular network. To reduce bandwidth requirements, the client compresses speech parameters sent to the server, and thus requires approximately 2.5-4.8 kbps of network bandwidth. A wireless local area network (LAN), which is currently used to simulate a wireless 3G network, connects the client to a Windows 2000 machine where CSR and SLU are performed. The client requires approximately 450 KB of code space and an additional 200 KB of runtime heap, and utilizes approximately 35% of the iPaq's 206 MHz StrongARM processor. At 2.5-4.8 kbps, we observed less than 5% relative error increase for the CSR engine. MiPad applications communicate via our dialog manager to both the CSR and SLU engines for coordinated context-sensitive *Tap and Talk* interaction, as shown in Figure 2.

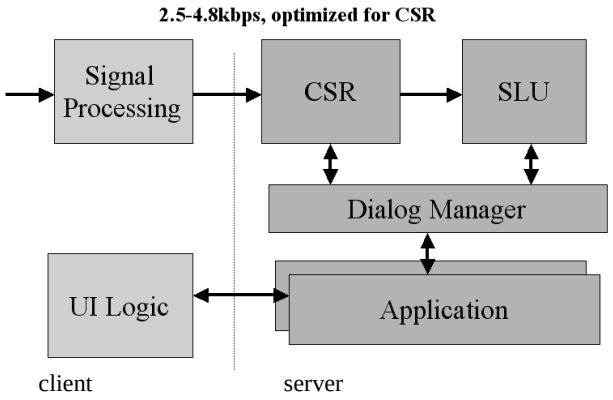


Figure 2 MiPad's client-server architecture. The client is based on a Windows CE iPaq, and the server is based on a Windows 2000 machine. The client-server communication is currently based on the wireless LAN.

2. MIPAD UI DESIGN

2.1 Tap and Talk interface

Because of MiPad's small form-factor, the present pen-based methods for getting text into a PDA (Graffiti, Jot, soft keyboard) are potential barriers to broad market acceptance. Speech is generally not as precise as mouse or pen to perform position-related operations. Speech interaction can also be adversely affected by the ambient noise. Moreover, speech interaction could be ambiguous without appropriate context information. 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 the user interface could increase by combining both.

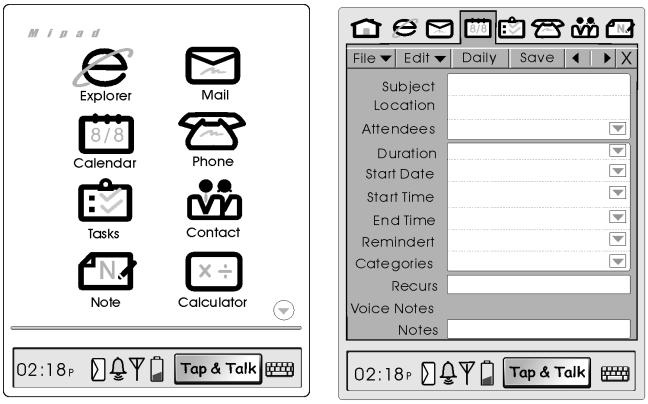
People tend to use speech to enter data and pen for corrections and pointing. As illustrated in Table 2, MiPad's *Tap and Talk* interface offers a number of benefits. MiPad has a *Tap & Talk* field that is always present on the screen as illustrated in MiPad's start page in Figure 3 (a) (the bottom gray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

Table 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

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 ASR.
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 the text about 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 commands by tapping the *Tap & Talk* field and talking to it. *Tap & Talk* avoids speech detection problems that are critical to noise environment deployment for MiPad. The appointment form shown on MiPad's display is similar to the underlying semantic objects. 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 slot-dependent language and semantic model.

2.2 Fuzzy soft keyboard

We can use the same n-gram in ASR to reduce the error rate of the soft keyboard. We model the position of the stylus tap as a

continuous variable, allowing the user to tap either in the intended key, or perhaps n earby 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.

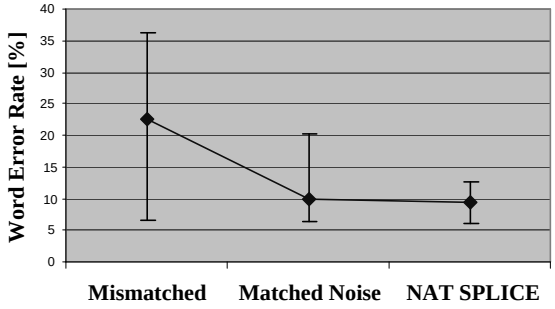


Figure 4 Word recognition error rates of close -talk microphone and built -in microphon e with or without noise adaptive training.

In the typical MiPad usage scenario, the user may use the built -in MiPad microphone that is very sensitive to environment noise. In a normal office environment, the word error rate on the WSJ dictation task differ s by a factor of two between the built -in microphone of Compaq's iPaq device, and a close -talk microphone. Since this error increase is mainly due to the additive environment noise, the Dr Who CSR engine used our noise adaptive training [1] to improve the performance of the built -in microphone.

Our noise robustness code has been improved to deal to improve the performance of the built -in microphone under both seen and unseen conditions [6,7]. For mismatched experiments, where noisy data was recognized with clean models, word error rates were as high as 36%. In matched experiments, a separate acoustic model was trained for each noise type and tested on similar data. This cut the average word error rate by better than half. Using NAT SPLICE, the average word error rate drops even more, and the maximum word error rate is reduced by over 1/3.

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"*. With n-gram framework, it is very expensive to capture long -span semantic information. The unified model uses a set of CFGs that captures the semantic structure of the domain. For the example listed here, we may have CFG's for <NAME> and <TIME> respectively, which can be derived from the factoid grammars. The training sentences now look like: *" Meeting <at three:TIME> with <Zhou Li:NAME>"*. and *"Meeting <at four PM:TI ME> with <Derek: NAME>*. With parsed training data, we can estimate the n -gram probabilities as usual. For example, *P(Zhou|three with)* is replaced by P(<NAME>|<TIME> 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 structure d classes. The structured class can be simple such as <DATE>, <TIME>, and <NAME> or can be complicated to contain 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-grams.

Most decoders can only support either CFGs or word n-grams. We have modified the decoder so that we can embed CFGs in the n -gram search framework to take advantage of the 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 either associated with a real-world entity or an action that the application takes on a real-entity. Each semantic class has slots that are linked with their corresponding CFG. In contrast to the sophisticated prompting response in voice -only conversational interface, the response is a direct graphic r endering of the semantic object on MiPad's display. After a semantic object g ets updated, the dialog manager fulfills the plan by executing both inter and intra -frame application logic and error repair strategy.

One of the critical tasks in SLU 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. For spontaneous

PIM application, Dr Who SLU engine's slot parsing error rate in the general *Tap and Talk* field is above 40%. 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 parsing error by more than 25%, which might still be insufficient to be useful. Fortunately, with our imposed context constraints in the *Tap and Talk* interface, where slot-specific language and semantic models can be leveraged, most of today's SLU technology limitations can be overcomed.

4. USER STUDY RESULTS

Our ultimate goal is to make MiPad produce real value to users. It is necessary to have a rigorous evaluation to measure the usability of the prototype. Our major concerns are "*Is the task completion time much better*?" and "*Is it easier to get the job done*?"

For our user studies, we set out to assess the performance of the current version of MiPad (with PIM features only) in terms of task-completion time and user satisfaction. 16 computer-savvy participants who had little experience with PDAs or speech recognition software used the partially implemented MiPad prototype. The tasks we evaluated include creating a new appointment and creating a new email. Each participant completed half the tasks using the tap and talk interface and half the tasks using the regular pen-only iPaq interface. We carefully counter-balanced the ordering of tap and talk and pen-only tasks.

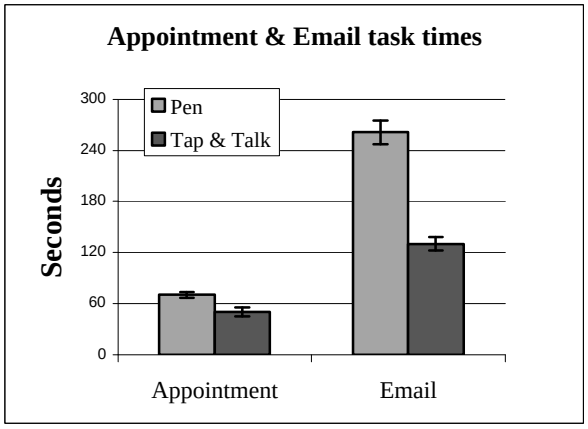


Figure 5 Task completion time of email transcription between the pen-only interface and *Tap and Talk* interface. The standard deviation is also shown above the bar of each performed task.

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, 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. On average it took the participants 50 seconds to create a new appointment with the tap and talk interface and 70 seconds with the pen-only interface. This is statistically significant, $t(15) = 3.29$, $p < .001$. The saving of time is about 30%. For transcribing an email it took 2 minutes and 10 seconds with tap and talk and 4 minutes and 21 seconds with pen-only. This difference is also statistically significant, $t(15) = 8.17$, $p < .001$. The saving of time is about 50%. Error correction for the *Tap and Talk* interface remains as one of the most unsatisfactory features. In our user studies, 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? - 15 of the 16 participants stated that they preferred using the tap and talk interface for creating new appointments and all 16 said they preferred it for writing longer emails. The preference data is consistent with the task completion times. Error correction for the Tap and Talk interface remains as one of the most unsatisfactory features. On a 7 point Likert scale, with 1 being disagree and 7 being agree, participants responded with a 4.75 that it was easy to recover from mistakes.

5. SUMMARY

MiPad is a work in progress for us to develop a consistent Dr Who interaction model and engine technologies for multimodal applications. Our currently application includes 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 voice-only telephony interaction.

The success of MiPad depends on spoken language technology and always-on wireless connection. With upcoming 3G wireless deployments in sight, the critical challenge for MiPad remains the accuracy and efficiency of our spoken language systems since it is likely MiPad may be used in a 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, and X. Lu, 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. "Robust Spoken Language Understanding in MIPAD". *ICASSP*-2001, Salt Lake City, UT

[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.

[6] L. Deng, et al. "*High-Performance Robust Speech Recognition Using Stereo Training Data*". in *ICASSP*-2001, Salt Lake City, UT

[7] J. Droppo, L. Deng and A. Acero. "*Efficient On-Line Acoustic Environment Estimation for FCDCN in a Continuous Speech Recognition System*". in *ICASSP*-2001, Salt Lake City, UT