A Generic Multimodal Architecture for Integrating Voice and Ink XML Formats

Zouheir Trabelsi
College of Telecommunication, The University of Tunisia, Tunisia

Abstract: *The acceptance of a standard VoiceXML format has facilitated the development of voice applications, and we anticipate a similar facilitation of pen application development upon the acceptance of a standard InkXML format. In this paper we present a multimodal interface architecture that combines standardized voice and ink formats to facilitate the creation of robust and efficient multimodal systems, particularly for noisy mobile environments. The platform provides a Web interactive system for generic multimodal application development. By providing mutual disambiguation of input signals and superior error handling this architecture should broaden the spectrum of users to the general population, including permanently and temporarily disabled users. Integration of VoiceXML and InkXML provides a standard data format to facilitate Web based development and content delivery. Diverse applications ranging from complex data entry and text editing applications to Web transactions can be implemented on this system, and we present a prototype platform and sample dialogues.*

Keywords: *Multimodal voice/ink applications, speech recognition, online handwriting recognition, mutual disambiguation, VoiceXML, InkXML.*

Received March 13, 2003; accepted August 17, 2003

1. Introduction

Although speech and pen technologies have improved significantly over the last decade, unimodal applications employing these technologies have been successful in only limited domains. However, the acceptance of a standard VoiceXML format has facilitated the development of generic, domain-specific voice applications, and many have been easily developed and are now widespread. We anticipate a similar facilitation of pen application development upon the further development and acceptance of a standard InkXML format. Also lacking at this time is a standard platform to facilitate the creation of multimodal applications by application developers. With the rapid spread of mobile phone devices and the convergence of the phone and the PDA, there is increasing demand for such a multimodal platform that combines the modalities of speech and pen to reach a greater population of users.

Because applications have generally become more complex, a limited multimodal system's architecture does not permit the end user to interact effectively across all tasks and environments [8]. Therefore, a major goal of multimodal system design is to support flexible and robust performance, even in noisy mobile environments. A multimodal interface should offer the user the freedom to use a combination of modalities, or to switch to a more suitable modality, depending on the nature of the task or environment. Thus, the system should be able to process parallel input from several modalities, and to integrate them to produce a semantically compatible overall interpretation. Since individual input modalities are suitable in some situations, and less ideal or even inappropriate in others, modality choice is an important design issue in a multimodal interface. Multimodal interfaces are expected to support a wider range of diverse applications, to accommodate a broader range of users than traditional unimodal interfaces, including users of different ages, skill levels, native language status, cognitive styles, sensory impairments, and other temporary or permanent handicaps or illnesses, and to cover a wider range of changing environmental circumstances.

We propose a multimodal architecture that combines VoiceXML and InkXML to develop multimodal voice/ink mobile applications for man-machine communication. This robust multimodal interface architecture should broaden the spectrum of users to the general population. Integration of VoiceXML and InkXML provides a standard data format to facilitate Web based development and content delivery. Diverse applications ranging from complex data entry and text editing applications to Web transactions can be implemented on this system. The architecture handles unimodal voice, ink, and touch-tone input, as well as combined multimodal voice/ink/touch-tone input. In addition, it employs mutual disambiguation of two or three input signals where each mode provides partial information and

dialogue context that aids in the interpretation of the other modes.

The paper is organized into the following sections: the strengths of the various modalities, a general multimodal architecture, the proposed platform, the current implementation and evaluation, and conclusions.

2. Strengths of Speech, Pen, and Touch-Tone Input

Speech offers speedy input and relative ease of use, and permits the user's hands and eyes to be simultaneously busy with a task, which is particularly valuable when users are in motion or in natural field settings. Users usually prefer speech for entering descriptive information and issuing commands. However, for some types of input, such as mailing addresses, current speech recognition engines do not provide sufficient speaker-independent, continuous speech recognition accuracy rates. Speech recognizers also have high error rates and a low tolerance for regional accents and other variations in speech (e.g., when the speaker has a cold), especially when the speaker's accent is outside the range used to train the system. In many of these situations handwriting systems can be more accurate.

Although the pen can be used to write words that are analogous to speech, it also can be used to convey symbols and signs, gestures, simple graphics, and to render signatures. In addition, it can be used to point, and to select visible objects like the mouse does in a direct manipulation interface. Pen input provides a more private and socially acceptable form of input in public settings, and a viable alternative to speech under circumstances of extreme noise. Thus, the pen offers critical capabilities for interacting with any form of graphic application, and it potentially can provide a very versatile and opportune base system, especially for mobile task.

Touch-tone digit recognition provides high accuracy in all environments, especially in noisy mobile ones. However, touch-tone input is limited and allows the user to enter only digits. Speech input can replace touch-tone digit input, but is usually less accurate. Unimodal applications that accept only touch-tone input are extremely limited and do not broaden the spectrum of the users. By accepting touch-tone input, a multimodal application offers the user an additional input modality so that the user can use it or switch to it depending on the specifics of the task or environment.

Combining speech, pen, and touch-tone inputs permits users to engage in more powerfully expressive and transparent dialogues, with speech and ink providing complementary capabilities [13] and touch-tone input providing a highly accurate fallback mode.

3. A General Multimodal Voice/Ink Architecture

This section discusses a general architecture for interpreting multimodal speech, Ink, and touch-tone digit input in a robust manner for noisy mobile environments. Many early multimodal systems that handle combined speech and pen input were able to process just speech combined with pen-based pointing input in a synchronized way [2]. More recent work shows that less than 20% of all users multimodal commands are of this limited type [15, 19]. Some multimodal systems do not support recognition of simultaneous modes, but only the recognition of alternative individual modes [9]. In addition, they do not allowed the user to freely use a modality or a combination of modalities, and to switch to a better-suited modality. Therefore, in noisy mobile environments the efficiency of such systems is poor due to low recognition rates and limited user freedom of modality. In addition, most recent multimodal systems [1, 5, 6, 9, 12] are based on limited architectures that do not support robust application development for noisy mobile environment or for a large spectrum of users. Also, such architectures limit the development of multimodal applications because they are not based on standard data formats, such as VoiceXML and InkXML, standard meaning representation structures, and general multimodal natural language processing. Few multimodal architectures [5] support mutual disambiguation of input signals to enable recovery from unimodal recognition errors. The proposed architecture supports the development of unconstrained multimodal applications that can handle speech, ink, and touch-tone input.

3.1. Semantic-Level Integration and Mutual Disambiguation of Complementary Modalities

There are two main architectures for multimodal systems. The first integrates signals at the feature level ("early fusion" or "feature fusion"). The second integrates information at a semantic level ("late fusion" or "semantic fusion") -- that is, the integration of the recognition outputs from both the speech recognizer and the handwriting recognizer. Systems that utilize the early feature-level approach generally are based on multiple Hidden Markov Models or temporal neural networks [3, 22] and the recognition process in one mode influences the course of recognition in the other. Feature fusion generally is considered more appropriate for closely coupled and synchronized modalities, such as speech and lip movements, for which both input channels provide corresponding information about the same articulated phonemes and words. However, such systems tend not to apply or

generalize well if they merge modes that differ substantially in the information content or time scale characteristics of their features. This is the case with speech and pen input for which the input modes provide different but complementary information that is typically integrated at the utterance level. In addition, modeling complexity, computational intensity, and training difficulty are typical problems associated with the feature-level integration approach, and large amounts of training data are required. We use the semantic-level approach [2, 4, 5, 23] that utilizes individual recognizers and a multimodal integration process. The individual recognizers can be trained using unimodal data, which are easier to collect and already publicly available for modalities like speech and handwriting.

A robust and well-designed multimodal system should be able to integrate complementary modalities such that the strengths of each modality are capitalized upon and used to overcome weaknesses in the other [13]. This general approach can result in a highly functional and reliable system. Mutual disambiguation involves recovery from unimodal recognition errors within a multimodal architecture, where semantic information from each input mode supplies partial disambiguation of the other mode. A well-integrated multimodal system can yield significant levels of mutual disambiguation between input signals, with speech disambiguating the meaning of ink and vise versa. Mutual disambiguation generates higher overall recognition rates and more stable system functioning than is possible by either individual technology [17].

3.2. Modality Choice and Flexible Input Processing

A multimodal interface should offer the user freedom to use a combination of modalities, or to switch to a better-suited modality, depending on the specifics of the task or environment. The system should be able to process parallel input from many modalities, and to integrate them and produce a semantically compatible overall interpretation. Since individual input modalities are well suited in some situations, and less ideal or even inappropriate in others, modality choice is an important design issue in a multimodal interface.

The system can interpret simultaneous speech, ink, touch-tone input. For example, the user can say "My name is" while writing his/her name on the pen tablet. To ensure proper synchronization, the system time stamps speech, pen, and touch-tone input events, and integrates them to form a frame-based description of the user's input. The speech thread generates events in response to output from the speech recognizer, and the pen thread generates selection events that are stored in a time-sorted buffer where they can be retrieved by the integrator thread. The integrator looks for a pen or touch-tone input that occurs closest in time to the spoken input. This design allows for asynchronous processing of multimodal input, and keeps pace with user input, while still processing them as coordinated multimodal pieces.

3.3. Error Handling and Late Confirmation

Compared with unimodal applications, a particularly advantageous feature of this multimodal application design is its ability to support superior error handling, both in terms of error avoidance and graceful recovery from errors. This multimodal application facilitates error recovery for both user-centered and application-centered reasons. First, empirical studies have demonstrated that users select the input mode (speech, ink, or touch-tone input) they judge to be less error prone, and this leads to fewer errors. Second, a user's language is simplified when interacting multimodally, which reduces the complexity of the natural language processing and thereby further reduces recognition errors. Third, users have a strong tendency to switch modes following systems errors, which facilitates error recovery. Finally, users report less frustration with errors when interacting multimodally, even when errors are as frequent as in the unimodal speech-only application. To take full advantage of such error handling, the speech and pen modes must provide parallel or duplicate functionality, meaning that users can accomplish their goals using either mode. Superior error handling results also from the mutual disambiguation process that allows users to recover from unimodal recognition errors, mainly from the speech recognizer in the noisy mobile environment.

In multimodal systems, to assure common ground is achieved, miscommunication is avoided, and collaborative effort is reduced, system designers must determine when and how confirmations ought to be requested. There are two main strategies: early confirmation in which confirmation is performed for each modality and late confirmation in which it is performed after the modalities have been merged. Based on earlier work [14], we adopt the late confirmation strategy that reduces the time to perform tasks because misunderstanding is reduced, users can interact faster, and the dialogues go more rapidly and efficiently.

3.4. Multimodal Grammar

The grammar used by the multimodal integrator is a set of rules defining the possible syntax among the speech vocabulary, ink vocabulary and touch-tone digits. Each rule is a sequence of objects (parts of speech). The speech recognizer, the handwriting recognizer, or the touch-tone digits recognizer can generate a word. This is important information used mainly during error handling. For each rule there is a corresponding action that should be executed once the rule is selected.

Rule i : <Oject1> + <Object2> + ...+ <Object n>
Object i : { (word1, sources1), …, (wordn, sourcesn)}
where source: (speech, ink, touch-tone)

3.5. Multimodal XML Language

Multimodal systems are expected to interact with many input or/and output devices and applications. Therefore to develop such systems, application developers need a standard language that allows them to define their application and interactions properly with input/output devices and other applications. The VoiceXML language does not support interaction with pen tablets and handwriting recognizers [7]; and InkXML does not support interaction with speech devices, speech recognizers, speech synthesizers, and touch-tone recognizers [10]. That is, they do not include tags that support the interaction with those devices and applications. Therefore, we suggest extending Voice/InkXML by including new tags to create a multimodal XML language to facilitate and standardize the development of multimodal voice/ink/touch-tone applications.

4. The Proposed Multimodal Voice/Ink System Platform

Figure 1 shows the proposed multimodal voice/ink architecture. The user may interact with the multimodal application using speech, ink, or touch-tone input devices. A V*oice Board* connects the system to the *Public Switched Telephone Network (PSTN)* [7], and performs basic media processing, such as touch-tone detection, call control, audio compression and decompression, media player, and media recorder.

The basic system has the following processors:

1. An *Enhanced Media board* (e.g., *Dialogic's Antares*) for speech recognition and synthesis (text-to-speech).
2. A *Handwriting Recognizer* to recognize handwriting input.
3. A *Database* for grammars, vocabularies, templates, and data used by the speech recognition, speech synthesis, and handwriting recognition processors, and by the application itself.
4. The *Multimodal Voice/Ink Information Processing Manager* for the logic that handles and controls all incoming and outgoing information from and to the multimodal application, as well as the initialization and termination of a session application with the user.

A future system could have the following additional processors:

1. A *Natural Language Processor* to handle advanced language processing.
2. A *Hand Drawing Recognizer and Graphic Generator* for hand drawing recognition and graphic generation. With these components the database may also contain appropriate graphic templates.
3. An *Enhanced Phone Device,* with an integrated pen tablet, allows a user to interact with the multimodal application over the PSTN network with speech, ink, and touch-tone input.

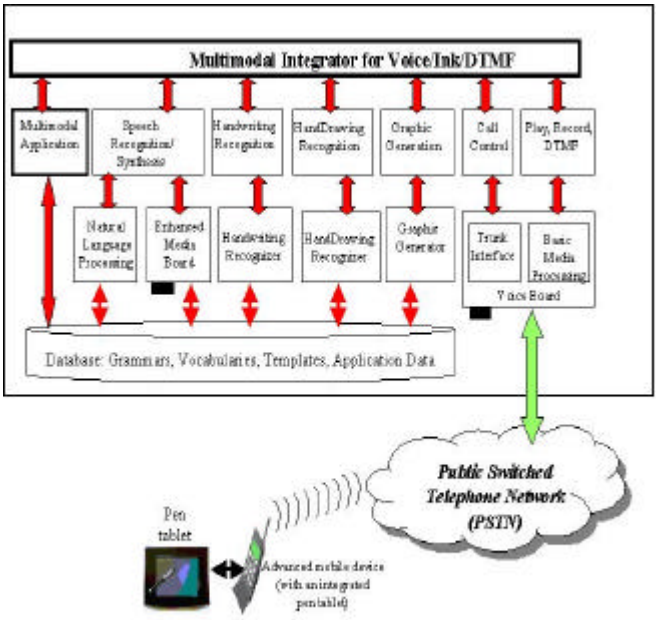

Figure 1. Multimodal voice/ink system platform.

4.1. Multimodal Voice/Ink Information Processing Flow

The system's architectural flow for processing multimodal input is illustrated in Figure 2. Speech, ink and touch-tone inputs are recognized in parallel by the speech recognizer, handwriting recognizer, and a touch-tone digit recognizer, respectively. The results from each recognizer are meaning fragment representations that are fused by the *Multimodal Integrator* to produce a semantically compatible unified interpretation. Here we summarize the responsibilities of each component, their interaction, and the results of their computation.

The *handwriting recognizer* recognizes the ink input. Recognition results consist of an n-best list (top n-ranked) of interpretations and associated probability estimates for each interpretation. The interpretation is encoded using a standard meaning representation structure, such as feature structures [11]. This list is then passed to the *Multimodal Integrator*.

The *Automatic Speech Recognizer (ASR)* offers a combination of relevant features: speaker-independent, continuous recognition, as well as multiple hypotheses and their probability estimates. The speech recognizer's output, like the handwriting recognizer's, is an n-best list of hypotheses and associated

probability estimates. These results are passed to the *Natural Language Processor (NLP)* for interpretation.

The *NLP* parses the output of the *ASR* to provide proper semantic interpretations. This process may introduce further ambiguity, that is, more hypotheses. Results of parsing are again in the form of n-best list. The results of the natural language processor are passed to the *Multimodal Integrator* for multimodal integration.

The *multimodal integrator* accepts feature/data structures from the handwriting recognizer, the natural language processor, and the touch-tone recognizer. The process of integration ensures that modes are combined according to a language specification, and that they meet certain multimodal timing constraints. These constraints place limits on when different input can occur, thus reducing error. Integrations that do not result in a completely specified command are ignored. The *multimodal integrator* then examines the joint probabilities for any remaining command and passes the feature structure with the highest joint probability to the *multimodal dialogue manager*. If no result exists, the feature structures are sent to the *mutual disambiguation processor* for possible error resolution. If no result exists, a message is sent to the user to inform him/her of the non-understandable input.

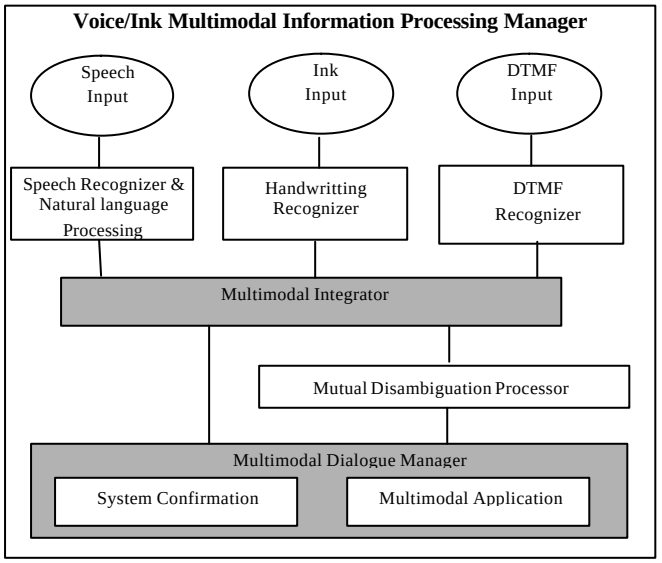

Figure 2. Multimodal voice/ink information processing.

5. Implementation

Based on the proposed multimodal architecture, a simple multimodal voice/ink system prototype has been developed. Figure 3 depicts the data flow in the system. The multimodal device is capable of handling voice/ink/touch-tone data, and displaying the feedback confirmation information on a small screen. The voice and ink data are processed by their respective voice and ink SDK (Software Development Kit). The output from the SDK is given to the multimodal integrator which consists of three main parts: the event handler which handles the different events generated by the voice and ink SDKs, the disambiguator which calculates the best results obtained from ink and voice, and the error handler/message passing which generates the confirmation message to be passed to the user and also handles errors generated with respect to the data and dialogue grammar. Each application developed on this system has its own set of grammar rules and dictionary. Therefore, by changing the grammar rules and dictionary we can use the system for various scenarios and applications.

The standardization of VoiceXML has simplified creation and delivery of Web-based, personalized interactive voice response services; enabled phone and voice access to integrated call-center databases, information and services on Web sites, and company Intranets; and helped enable new voice-capable devices and appliances. VoiceXML is designed for creating audio dialogues that feature synthesized speech, recording of spoken input, and the recognition of spoken and touch-tone (DTMF) input. A typical VoiceXML application consists of a dialogue template where the application requests minimal information required from the user in order to access the data. The left-hand side of the figure depicts the voice portion of the system. In a voice/touch-tone only application the user telephones the system and inputs voice or DTMF signals that are recognized, the system follows the VoiceXML dialogue to respond with synthesized or recorded speech, and interaction continues until termination of the application dialogue.

The currently proposed InkXML format contains only the raw ink data and the handwriting recognizer's output that corresponds to the ink data. InkXML documents are currently a medium to store ink data from various pen devices, and this facilitates the storage, manipulation, and exchange of large amounts of ink data in a common format. However, in its current format the InkXML does not have the capability to contain dialogue or to produce visual output. The right-hand side of the figure depicts the ink portion of the system, currently consisting of a Wacom pen tablet for input, non-InkXML format for the data, and standard techniques for producing visual output to the screen (XHTML is a often used for this purpose). For ink only applications interaction is initiated by the system producing visual output on the screen requesting a selection or input from the user, and interaction continues following the application dialogue until termination. We anticipate moving to an InkXML framework shortly, at least for the data format, and we encourage further development to extend InkXML to include dialogues and standardized formats for producing visual output.

One of the applications we have implemented is a banking application. When a user calls the system, the application prompts him to enter his bank account number, using one of the three available input modalities-speech, ink or touch-tone digits. If the bank

account is identified, then the application prompts the user to choose from a menu of options: account information, or personal information update. To update personal information, the user is asked to update any of a number of fields that he selects, such as *name*, *telephone number*, *address*, and *e-mail*. For the *name and telephone number* fields, the user may use one of the three available input modalities (speech, ink, or touch-tone digits) to enter his data. However, since the speech recognition engine used in this implementation has a high error rates in noisy mobile environment and low tolerance for speaker-independent, continuous speech recognition, the application recommends, but does not require, that the user use the Pen tablet to fill the *address* and *e-mail* fields.

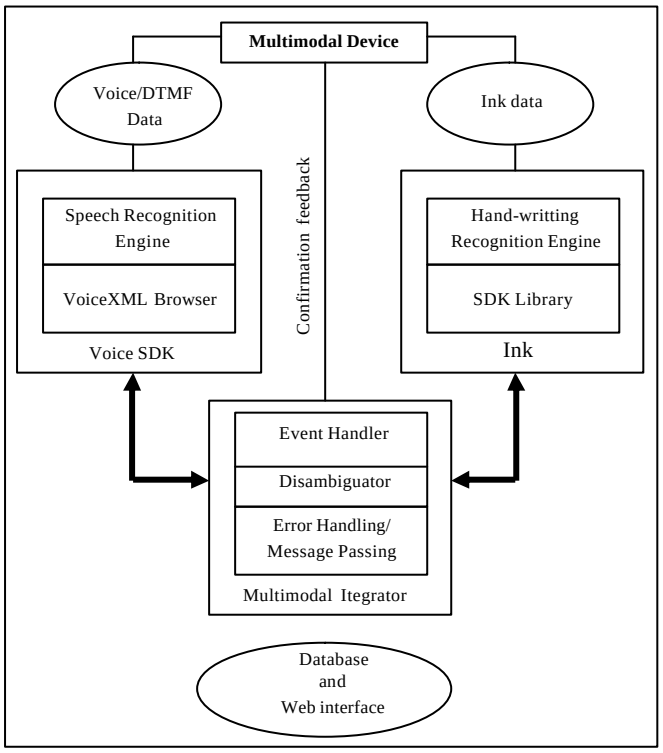

Figure 3. Multimodal voice/ink system prototype.

6. Other Applications

A prominent application, which highlights the need for a multimodal interface is tracking personal information (a multimodal diary). The multimodal diary contains all the personal information regarding an individual such as:

- Personal stocks.
- Banking information.
- Travel schedules tracking and reservation.
- Address history.
- Telephone diary.
- Appointment and special day reminder.

The above listed information has to be inputted and updated in the system. Updating information such as addresses, appointment details, bank information and stocks through voice interface can be tedious and time consuming. Alternatively we can use pen for inputting and updating the same. The error rate and time consumption are considerably lowered when we use ink over voice media.

On a similar note, while reviewing the information about stocks, bank, etc., its much easier to output with the voice media. Since voice offers the convenience of mobility to the user. The problem of switching from voice to ink and back is solved using the systems proactive role. Following conversation highlights the above solution.

(Application: appointment and special day reminder)
System: Say which operation would you like to perform: update, delete, or add?
User: Update
System: Did you say Update?
User: Yes
System: Please say the record number you want to update?
User: Record Five
System: Did you say five?
User: Yes
System: Which part would you like to update. Say text or time?
User: Text
System: Did you say text?
User: Yes
System: Please use ink to input the new text.
(Control passes to the ink medium. The system waits for the user to input the new text and to submit it and, upon submission, the control switches back to voice.)
(Confirmation step)
System: Your information for record five has been updated.

As illustrated in the above sample dialogue, the application relies heavily on the synchronization of the voice and ink media. We find that such applications require the system to be proactive and the user to be passive-that is, *the system decides the path that the user takes*. Additional sample dialogues are presented in the appendix.

7. Evaluation

Preliminary evaluation confirms that some input types are more appropriate for voice and others for pen. For example, names and addresses are more accurately captured by pen while numeric input and simple choices are easily and usually more conveniently handled by voice. We anticipate that more extensive evaluation of the completed system will show that the multimodal architecture provides more stable and robust applications, particularly in noisy environments.

Compared with speech-only interaction with the same application, preliminary empirical work with users demonstrated that multimodal pen/voice interaction resulted in faster task completion time,

fewer content errors, and fewer spontaneous disfluencies. The error handling process has been dramatically reduced. The dialogues between the users and the application went more rapidly and efficiently, increasing user satisfaction in comparison to the same speech-only application. Thus, the evaluation of this application shows that the combined use of different input modes (speech, ink, and touch-tone digits) leads to the development of more robust and friendly multimodal applications.

A particularly advantageous feature of this multimodal architecture is its ability to support error handling, compared with the same unimodal speech-only architecture, both in terms of error avoidance and graceful recovery from errors. The evaluation process has demonstrated that users select the input mode (speech, ink, touch-tone) that they judge to be less error prone for particular input, which leads to error avoidance. In addition, the users have a strong tendency to switch modes after systems errors to further facilitate error recovery. Since users can use three modalities, they report less frustration with errors when interacting multimodally, even when errors are as frequent as in the unimodal speech-only application. Mutual disambiguation has contributed considerably in many occasions to recover from unimodal recognition errors, mainly from the speech recognizer.

8. Conclusion

Our proposed architecture for developing multimodal voice/ink applications for noisy mobile environments combines different input modalities to facilitate the development of robust and friendly multimodal applications supporting superior error handling. Our interactive system platform should assist developers in the creation of generic multimodal applications.

We envision that users will soon employ smart devices such as wireless phones with integrated pen tablets and more powerful processing capabilities to take full advantage of the proposed multimodal voice/ink architecture. Such smart devices should be able to perform locally enhanced media processing, such as voice recognition, speech synthesis, and handwriting recognition. Graphic generation capabilities on the user's pen tablets should also enhance the efficiency of multimodal applications and may allow for the development of applications for a broader spectrum of the population, including permanently and temporarily disabled users. Finally, we suggest that a specialized XML be developed to facilitate the development of multimodal applications.

Appendix

Application: Banking Information
System: You can access your existing account or you can open a new account. What would you like to do?
User: Check existing account
System: Did you say existing account?
User: Yes
System: Please enter your account number by ink.
(Control passes to the ink media. The system waits for the user to input the new text and submit. Once the user has submitted the data the control switches back to voice.)

System: Did you write one eight seven four six five?
User: No
System: Please enter your account number by ink.
(Control passes to the ink media. The system waits for the user to input the new text and submit. Once the user has submitted the data the control switches back to voice.)

System: Did you write one eight one four six five?
User: Yes
System: Please speak your four digit, pin number
User: one two three four
System: Did you say one two three four?
User: Yes
System: Choose personal information, checking or savings.
User: personal information
System: Did you say personal information?
User: Yes
System: What would you like to do? Access your information or change your information.
User: Change information
System: Did you say change information?
User: Yes
System: Would you like to change the address or telephone number or exit?
User: Address
System: Did you say address?
User: Yes
System: Please enter your new address by ink
(Control passes to the ink media. The system waits for the user to input the new text and submit. Once the user has submitted the data the control switches back to voice.)
System: Did you write *one martine av white plains new york one zero six zero three*?
User: Yes
System: Your address has been changed.

Application: Telephone Diary
System: Would you like to change the telephone number or access telephone numbers or say exit?
User: change telephone number
System: Did you say change telephone number?
User: Yes
System: Please say the name of the person or contact whose telephone number you would like to change.
User: Thomas
System: Sorry I couldn't understand or may be the name is not in your diary. Please speak the name again

User: Tommy
System: Did you say Tommy?
User: Yes
System: Please enter the new telephone number for Tommy by ink
(Control passes to the ink media. The system waits for the user to input the new text and submit. Once the user has submitted the data the control switches back to voice.)
System: Did you write nine one four nine six four zero eight four one?
User: No
System: Please enter again your new telephone number by ink
(Control passes to the ink media. The system waits for the user to input the new text and submit. Once the user has submitted the data the control switches back to voice.)
System: Did you write nine one four nine five four zero eight four one?
User: Yes
System: Your telephone number has been update.

References

[1] Bers J., Miller S., and Makhoul J., "Designing Conversational Interfaces with Multimodal Interaction," *DARPA Workshop on Broadcast News Understanding Systems*, pp. 319-321, 1998.

[2] Bolt R. A., "Put-that-three: Voice and Gesture at the Graphics Interface," *Computer Graphics*, vol. 14, no. 3, pp. 262-270, 1980.

[3] Bregler C., Manke S., Hild H., and Waibel A., "Improving Connected Letter Recognition by Lip Reading," *in Proceedings of Int. Conference Acoustics, Speech and Signal Processing*, IEEE Press, vol. 1, pp. 557-560, 1993.

[4] Codella C., Jalili R., Koved L., Lewis J., Ling D., Lipscomb J., Rabenhorst D., Wang C., Norton A., Sweeney P., and Turk C., "Interactive Simulation in a Multi-Person Virtual World," *in Proceedings of Conference on Human Factors in Computing Systems (CHI'92)*, ACM Press, New York , pp. 329-334, 1992.

[5] Cohen P. R., Johnston M., McGee D., Oviatt S., Pittman J., Smith I., Chen L., and Clow J., "Quickset: Multimodal Interaction for Distributed Applications," *in Proceedings of Fifth ACM Int. Multimedia Conference*, ACM Press, New York, pp. 31-40, 1997.

[6] Duncan L., Brown W., Esposito C., Holmback H., and Xue P., "Enhancing Virtual Maintenance Environments with Speech Understanding," *Boeing M and CT TechNet*, 1999.

[7] Edgar B., *The VoiceXML Handbook*, CMP Books, 2001.

[8] Fujisaki T., Modlin W., Mohiuddin M. K., and Takahashi H., "Hybrid On-Line Handwriting Recognition and Optical Character Recognition System," *U.S. Patent* 6.011.865, 2000.

[9] Holzman T. G., "Computer Human Interface Solutions for Emergency Medical Care," *Interactions*, vol. 6, no. 3, pp. 13-24, 1999.

[10] InkXML Documents, http://www.easystreet.com/~lartech/InkXML/.

[11] Kay M., "Functional Grammar," *in Proceedings of Fifth Annual Meeting of the Berkeley Linguistics Society*, pp. 142-158, 1979.

[12] Lai J. and Vergo J., "MedSpeak: Report Creation with Continuous Speech Recognition," *in Proceedings of Conference on Human Factors in Computing (CHI'97)*, ACM Press, pp. 431-438, 1997.

[13] Larson J. A., Oviatt S. L., and Ferro D., "Designing The User Interface for Pen and Speech Applications," *in Proceedings of Conference on Human Factors in Computing Systems (CHI'99)*, Philadelphia, PA, 1999.

[14] McGee D., Cohen P. R., and Oviatt S. L., "Confirmation in Multimodal Systems," *in Proceedings of Int. Joint Conference of Association for Computational Linguistics and the International Committee on Computational Linguistics (COLING-ACL'98)*, University of Montreal Press, pp. 823-829, 1998.

[15] McNeill D., *Hand and Mind: What Gestures Reveal about Thought*, University of Chicago Press, Chicago, 1962.

[16] Oviatt S. L., "Multimodal Interactive Maps: Designing for Human Performance," *Human-Computer Interaction (special issue on Multimodal Interfaces)*, vol. 12, pp. 93-129, 1997.

[17] Oviatt S. L., "Mutual Disambiguation of Recognition Errors in Multimodal Architecture," *in Proceedings of Conference Human Factors in Computing Systems (CHI'99)*, ACM Press, New York, pp. 576-583, 1999.

[18] Oviatt S. L., "Pen/Voice: Complementary Multimodal Communication," *in Proceedings of Speech Technology*, New York, 1992.

[19] Oviatt S. L. and Van G R., "Error Resolution During Multimodal Human-Computer Interaction," *in Proceedings of Int. Conference on Spoken Language Processing*, University of Delaware Press, pp. 204-207, 1996.

[20] Oviatt S. L., Cohen P. R., Wu L., Vergo J., Duncan L., Suhm B., Bers J., Holzman T., Winograd T., Landay J., Larson J., and Ferro D., "Designing the User Interface for Multimodal Speech and Pen-Based Gesture Applications: State-of-the-Art Systems and Future Research Directions," *Human Computer Interaction*, vol. 15, no. 4, pp. 263-322, 2000.

[21] Oviatt S. L., DeAngeli A., and Kuhn K., "Integration and Synchronization of Input Modes During Multimodal Human-Computer Interaction," *in Proceedings of Conference on Human Factors in Computing Systems (CHI'97)*, New York, pp. 415-422, 1997.

[22] Pavlovic V., and Huang T. S., "Multimodal Prediction and Classification on Audio-Visual Features," *AAAI'98 Workshop on Representations for Multi-modal Human-Computer Interaction*, AAAI Press, Menlo Park, CA, pp. 55-59, 1998.

[23] Wang J., "Integration of Eye-Gaze, Voice and Manual Response in Multimodal User Interfaces," *in Proceedings of IEEE Int. Conference Systems, Man and Cybernetics*, IEEE Press, pp. 3938-3942, 1995.

Zouheir Trabelsi received his PhD from Tokyo University of Technology and Agriculture, Japan, in the field of computer science, March 1994. From April 1994 until December 1998, he was a computer science researcher at the Central Research Laboratory of Hitachi in Tokyo, Japan. From November 2001 until October 2002, he was a visiting assistant professor at Pace University, New York, USA. Currently, he is an associate professor at the College of Telecommunications, the University of Tunisia. His research areas are mainly multimodal voice and ink systems, human computer interaction, internet/networking hacking and security and the TCP/IP protocols.