

Robust speech interaction in a mobile environment through the use of multiple and different media input types

Rainer Wasinger, Christoph Stahl, Antonio Krueger

Department of Intelligent User Interfaces

DFKI, GmbH

(`rainer.wasinger`, `cristoph.stahl`)@dfki.de, `krueger@cs.uni-sb.de`

ABSTRACT

Mobile and outdoor environments have long been out of reach for speech engines due to the performance limitations that were associated with portable devices, and the difficulties of processing speech in high-noise areas. This paper outlines an architecture for attaining robust speech recognition rates in a mobile pedestrian indoor/outdoor navigation environment, through the use of a media fusion knowledge component.

1. Introduction

We are now seeing technological advances in mobile devices in the areas of speed, memory, physical size, encompassed features, as too the reliability in wireless connection protocols such as infra-red, bluetooth and wireless LANs. Users no longer wish to be restricted to the office or desktop computer for human-computer-interaction, but instead desire the aid of computing devices, inside, outside, and on the go, yet with the same quality performance available on the desktop. Such a modern environment requires intuitive and robust interaction techniques, between the user, the embedded device(s) and the surrounding environment.

Through the fusing of different user-inputs such as speech and gesture, and the further enriching of this information through sensor data, an extremely robust and flexible interface can be created for mobile devices, in an environment containing ever-changing user-contexts and high levels of noise. This paper outlines such an architecture for pedestrian navigation.

2. Related Work

There are many examples of multimodal systems that use speech and keyboard/mouse inputs, but these often do not fuse the different inputs together. Instead, they allow the user to use either one, or the other modality at a time. These systems are even less common on mobile devices and in difficult environments like the outdoors.

Two systems that resemble parts of our system are that of DFKI's Smartkom [1], and SRI's QuickSet [2]. Smartkom is a vast project that spans intuitive multi-modal communication, for the home/office, public, and mobile scenarios. SRI's QuickSet project is also multi-modal, based on a pen/voice system running on a handheld PC, and

communicating via wireless LAN through a distributed agent architecture.

Unlike these and similar projects, all of the recognition engines (speech, gesture) are now embedded on the PocketPC itself and designed for the outdoor environment. Rather than the processing be carried out remotely, or even in a distributed fashion, it now all takes place in real time on the PDA. Furthering this, it is not only the information from the recognition engines that is used in the media fusion, but also information from sensors connected to the PDA. This is an important feature of our system.

A final distinction is that our embedded architecture takes both intra- and extra-gestures into consideration. This means that the user, aside from interacting with items on their PDA screen, can now also interact with objects in the real world by pointing at them. This form of interaction closely resembles user-user interaction, rather than the common user-PDA interaction that was seen up until now.

3. Pedestrian Navigation Scenario

The pedestrian navigation scenario allows the user to plan one or more routes over the Internet, and download these to their PDA. These routes can be either indoor or outdoor. Upon downloading the routes, the user can select the one they are currently interested in, and be directed along the route through the use of speech and graphics. Other functions of the navigation system include a birdseye and egocentric mode (birdseye is shown in Figure 1), zoom capabilities, a memo feature that stores notes (text / graphics / speech) to the user's localised coordinate on the map, and the ability to re-listen to position-sensitive information regarding remaining distance to the next street segment and nearby landmarks.

Aside from the navigation mode, the user can at any time switch into an exploration mode. In this mode they can leave their planned route and explore their surrounding environment, querying what it is that they see. Often these user-queries take the form of combined speech and gesture, in which the user can speak and point simultaneously at objects on the PDA display in front of them, or alternatively point at objects in the real world while speaking to the system. Common user-requests (currently in German) take the form of: "what is that?" and "describe this landmark to

me". The interface as seen by the user is shown in Figure 1 below.

User: How do I get from here to here?



Figure 1: Pedestrian navigation interface.

Finally, the user can load up old routes to re-trace their steps and analyse the notes they have made. This creates an effective simulation environment, often required for demonstration purposes.

4. User Interactions

4.1. Intra and extra gestures

The interaction channels for the user are that of speech and gesture. Gesture interaction takes two forms, that of intra-gesture (pointing to objects on the PDA's screen), and extra-gesture (pointing to objects in the real world). Intra-gestures are performed using the PDA's stylus. Extra-gestures require sensor information from the PDA to function correctly, such as the user's direction and speed originating from the PDA's compass and GPS units respectively. With this available information, the user can point to objects in the distance while speaking and have the system understand what is being referred to.

4.2. Speech and speech-gesture combined interaction

There are three different interaction methods required for complete multi-modal interaction using speech and gesture. These are speech-only, gesture-only, and speech and gesture combined. This paper is only concerned with the differences between speech-only, and fused speech-gesture interactions.

Speech-only interactions include for example: "tell me about the Mensa" or "how do I get from the Informatik building to the Uni-film building?" A user who is not aware of their task must however already be familiar with the layout of the area in order to know the object names, or the information must be presented to them on the screen. This requires careful consideration as to what and how much information can be displayed to the user on the small PDA display. Along with half a dozen landmarks (parks, monuments, shopping centres, buildings such as post-offices and schools, traffic lights, zebra crossings), there are also street names that

require labeling. It soon becomes clear that one cannot simply label everything on the screen. For a user who is aware of their task, speech overcomes this by allowing the user to extract only the information required. Alternative methods in presenting information would be to implement speech lists, different zoom levels, or filters displaying only certain types of landmarks. Our system uses a hybrid between some of these interaction methods and fused speech-gesture input interactions.

Fused speech-gesture input interactions include for example: "What is this?", "take me to here", "how long would it take me to get from here to here?" and "tell me about that church over there". It can be seen in all of these examples that a 'wild-card' is used to refer to something that the user can see, but may not know how to pronounce. This increased flexibility, and the robustness in speech recognition that it provides are discussed below.

4.3. Increased flexibility and robust recognition rates

The first advantage of fusing inputs, aside from saving valuable screen space on the PocketPC, is that the user's interaction is given an extended degree of flexibility [3]. Users quite often have difficulty asking for information, simply because they do not know how to create a syntactically complete question. In providing a combination speech-gesture input, the user is still able to represent the query with the underlying semantics intact.

The second advantage is the potential for more robust recognition rates. This is supported by tests on mutual disambiguation [3,4], and also by our own initial domain-specific tests. One common area of improvement is in noun resolution, for example when the user points to the Mensa building while saying "tell me about the Mensa building over there". This example provides two non-conflicting information sources for the noun, and can therefore be more confidently resolved.

5. Media Fusion Architecture

5.1. Technology / Software

The target implementation environment consists of a PocketPC running the WinCE operating system, and overlaid by the PocketPC 2002 platform. The PDA has inbuilt IRDA support to receive information from the infrared beacons, and bluetooth capabilities used for communicating with the GSM/UMTS mobile phone and GPS device. An alternative CF GPS device from Insense contains a compass, which together with the GPS provides speed, acceleration, direction and position information.

The PDA obtains its map-information via a HTTP connection with a server containing the Geographic Information System (GIS). This PDA-Server connection can be either USB (to a desktop) or bluetooth (over mobile phone).

Source Type	MI	MI Type	Unique ID	Time	Conf. Value	Modality Ptr. or Value
PDA	Sensor	Direction	100	1046788415000	--	directionVal
PDA	Sensor	Velocity	101	1046788415000	--	velocityVal
PDA	Speech	Microphone	102	1046788415000	0.8	*speechPtr
PDA	Gesture	Stylus	103	1046788415500	0.9	*objectPtr

Table 1: Inputs as written to the media fusion blackboard.

The main supporting software comprises ParallelGraphics' Cortona VRML [5] viewer for the interactive 3D graphics and IBM's Embedded ViaVoice for the speech engine. This speech engine consists of a formant-based synthesizer and a dynamic unlimited vocabulary recognizer. Although both the German and English ASR/TTS binary data files exist on the system, the main development is being carried out in German.

5.2. Media fusion inputs

The fusion of different media inputs to form a single modality-free and unambiguous representation is known as media fusion. As described in [6], inputs are made up of the 'modalities', the 'code' and the 'media'. In our system, the modalities correspond to the auditory and visual senses. The code refers to the communication symbols required by each modality, such as language, graphics and gesture. The media refers to the actual communication channel, for example stylus, pointing, or microphone.

The user's speech is converted to language, while the intra- and extra-gestures are converted to graphical objects or world-model coordinates. Extra-gestures are based on direction data obtained from the sensors, and assumptions made on how far away the object that the user is pointing to really is.

5.2.1. Speech recognition

Speech recognition is made possible through the use of a set of grammars that include both the interaction commands and the names of the objects that the user can interact with. These grammars are kept fairly small (around 100 words), and relate specifically to the currently loaded map. This is in contrast to grammars that cover street navigation for the whole of a country, which would undoubtedly place a burden on the PocketPC.

Although the returned results are currently simple word-lattices, we expect to integrate a language understanding module called SPIN [7] that will transform these word-lattices into a deeper semantic representation consisting of nested frames / typed feature structures. This has already been ported to the PDA, but still requires work in interfacing the Java classes from our C++ code.

5.2.2. Gesture recognition

Two types of intra-gestures are currently recognized by our system. The first is a pointing gesture used to query landmarks. The second is a slide gesture used to query street names. These are shown in Figure 2. When the user touches a landmark, an object reference is returned by the Cortona

VRML viewer, otherwise map-coordinates (longitude, latitude) are returned. Important to note, is that all gestures are mapped from the 2D screen to the 3D scene space. Simple program control and command gestures are not considered. Extra-gestures function similarly, but have coordinates calculated through the location and orientation of the PDA in the 3D-plane.



Figure 2: Point and slide gestures, used to query landmark and street names.

5.3. Blackboard design

The media fusion architecture is based on a blackboard design, in which data from each of the embedded recognizers and sensors is recorded for further analysis by the media fusion knowledge component. Events of type speech and gesture are triggered by the user interacting with the system. Sensor data which is constantly changing is first screened by the blackboard manager prior to storing, so that only significant changes are recorded to the blackboard. A set of possible entries to the blackboard can be seen in Table 1. Each entry is categorized by type, and is given a timestamp indicating when the event occurred in milliseconds, a confidence value from the relevant recognizer, and either a value or a pointer to the actual object in question.

Currently only the best result is entered on the blackboard by each recognizer. This means that the individual recognizers are left in charge of determining the best result. It is foreseeable in the future that the n-best results be passed on to the media fusion component, which then determines the best fit by taking additional factors into account, such as the user's history.

Figure 3 illustrates the interaction of the media fusion component with the individual recognition and data sources.

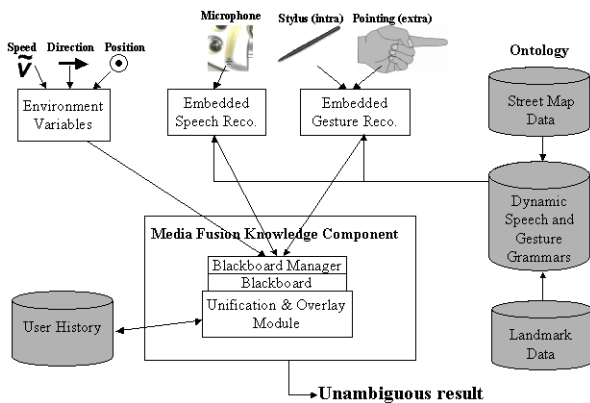


Figure 3: Data flow between the media fusion knowledge component and the recognition and data sources.

5.4. Overlay and unification

Until now, we have assumed that all inputs on the blackboard are non-conflicting. Described in [8], there are two basic operations for discourse processing in a multimodal dialogue system – unification [2] and overlay. This notion is also implemented in our work to solve the problem of conflicting information types and reference disambiguation.

When no conflicting information is present on the blackboard, the commutative operation of unification is used to combine the multiple inputs into one. However, as [8] describes, when conflicting information between the ‘covering’ and the ‘background’ exist, the process of overlay is required. This procedure is guaranteed to return a single result, which ultimately means that part of the conflicting information is overwritten. Aside from timestamps, our parameters in determining what information is relevant, are derived from the user’s history, the time interval of the user’s requests (generally around 2 - 8 seconds per user request), and probability tables outlining how likely a sequence of user-input types (speech and gesture) are to occur in the given context.

6. Discussion

Initial usage of the system during development has shown that the media fusion component consistently produced more robust recognition rates compared to speech alone. The two test environments considered for high levels of noise were that of the central university bus stop (outside), and the students cafeteria (inside). At this stage, a full user-study has not been conducted. The tests that we intend to conduct on our pedestrian navigation scenario include analysis on the proportion of time, and the type of tasks that people use speech, gesture and combined speech-gesture for. We also intend to calculate the improvement factor that occurs in our pedestrian specific domain, resulting from the fusion of multiple media inputs.

7. Future Work

The use of media fusion to create a means to robust speech recognition in mobile and outdoor environments is still fairly

new. Future work will take place on two fronts. The first front will take the form of integrating deeper levels of natural language processing, and semantic representation of the data in the embedded environment. The scoring techniques used in unification and overlay will also be adequately optimised, detailed user-studies will be performed, and the speech input/output will also be extended to cover English. The second front will develop other mobile scenario possibilities, in particular, shelf-selection incorporated in an intelligent environments scenario.

8. Conclusion

This paper has presented the implementation of a media fusion knowledge component in a pedestrian navigation system, to obtain robust recognition in noisy, mobile, indoor and outdoor environments. It has described the supporting media fusion architecture, the methods in which non-conflicting and conflicting information sources can be fused together, and the type of interactions common between the user, the PDA and the environment.

9. Acknowledgements

The pedestrian navigation system described above is being designed at the University of Saarland, together with the REAL project. The focus of this work falls under the project COLLATE (Computational Linguistic and Language Technology for Real Life Applications), in particular, the project M3I (Mobile Multi Modal Interaction).

10. References

- [1] Wahlster, W., Reithinger, N., Blocher, A. “SmartKom: Multimodal Communication with a Life-Like Character”, *Proc. of Eurospeech 2001*, pp. 1547-1550.
- [2] Cohen, P., Johnston, M., McGee, M., Oviatt, S., Pittman, J., Smith, I., Chen, L. & Clow, J. “Quickset: Multimodal interaction for distributed applications”, *Proceedings of the Fifth ACM International Multimedia Conference*, 1997, pp 31-40.
- [3] Oviatt, S., “Ten myths of multimodal interaction”, *Communications of the ACM*, v.42 n.11, 1999, pp 74-81.
- [4] Oviatt, S., “Mutual Disambiguation of Recognition Errors in a Multimodal Architecture”, 1999, pp. 576-583, *CHI*.
- [5] VRML 97: The Web 3D Consortium, URL: http://www.web3d.org/fs_specifications.htm.
- [6] Wahlster, W., “SmartKom: Multimodal Dialogs with Mobile Web Users”, *Proc. of the Cyber Assist International Symposium*, 2001, pp 33 - 34.
- [7] Engel, R., “SPIN: Language Understanding for spoken Dialogue systems using a production system approach”, *ICSLP*, 2002.
- [8] Alexandersson, J., Becker, T., “Overlay as the Basic Operation for Discourse Processing in a Multimodal Dialogue System”, *Proceedings of the IJCAI Workshop ‘Knowledge and Reasoning in Practical Dialogue Systems*, 2001.