An Adaptive Approach to Collecting Multimodal Input

Anurag Gupta
University of New South Wales
School of Computer Science and Engineering
Sydney, NSW 2052 Australia
akgu380@cse.unsw.edu.au

Abstract

Multimodal dialogue systems allow users to input information in multiple modalities. These systems can handle simultaneous or sequential composite multimodal input. Different coordination schemes require such systems to capture, collect and integrate user input in different modalities, and then respond to a joint interpretation. We performed a study to understand the variability of input in multimodal dialogue systems and to evaluate methods to perform the collection of input information. An enhancement in the form of incorporation of a dynamic time window to a multimodal input fusion module was proposed in the study. We found that the enhanced module provides superior temporal characteristics and robustness when compared to previous methods.

1 Introduction

A number of multimodal dialogue systems are being developed in the research community. A common component in these systems is a multimodal input fusion (MMIF) module which performs the functions of collecting the user input supplied in different modalities, determining when the user has finished providing input, fusing the collected information to create a joint interpretation and sending the joint interpretation to a dialogue manager for reasoning and further processing (Oviatt et. al., 2000). A general requirement of the MMIF module is to allow flexibility in the user input and to relax any restrictions on the use of available modalities except those imposed by the application itself. The flexibility and the multiple ways to coordinate multimodal inputs pose a problem in determining, within a short time period after the last input, that a user has completed his or her turn. A method, Dynamic Time Windows, is proposed to address this issue. Dynamic Time Windows allows the use of any modality, in any order and time, with very little delay in determining the end of a user turn.

2 Motivation

When providing composite multimodal input, i.e. input that needs to be interpreted or combined together for proper understanding, the user has flexibility in the timing of those multimodal inputs. Considering two inputs at a time, the user can input them either sequentially or simultaneously. A multimodal input may consist of more than two inputs, leading to a large number of composite schemes. MMIF needs to deal with these complex schemes and determine a suitable time when it is most unlikely to receive any further input and indicate the end of a user turn.

The determination of the end of a user turn becomes a problem because of the following two conflicting requirements:

1. For naturalness, the user should not be constrained by pre-defined interaction requirements, e.g. to speak within a specified time after touching the display. To allow this flexibility in the sequential interaction metaphor, the user can provide coordinated multimodal input anytime after providing input in some modality. Also each modality has a unique processing time require-

ment due to differing resource needs and capture times e.g. spoken input takes longer compared with touch. The MMIF needs to consider such delays before sending information to a dialogue manager (DM). These requirements tend to increase the time to wait for further information from input modalities.

2. Users would expect the system to respond as soon as they complete their input. Thus, the fusion module should take as little time as possible before sending the integrated information to the dialogue manager.

3 The MMIF module

We developed a multimodal input fusion module to perform a user study. The MMIF module is based on the model proposed by Gupta (2003). The MMIF receives semantic information in the form of typed feature structures (Carpenter, 1992) from the individual modalities. It combines typed feature structures received from different modalities during a complete turn using an extended unification algorithm (Gupta et. al., 2002). The output is a joint interpretation of the multimodal input that is sent to a DM that can perform reasoning and provide with suitable system replies.

3.1 End of turn prediction

Based on current approaches, the following methods were chosen to perform an analysis to determine a suitable method for predicting the end of a user turn:

1. Windowing - In this method, after receiving an input, the MMIF waits for a specified time for further input. After 3 seconds, the collected input is integrated and sent to the DM. This is similar to Johnston et. al. (2002) who uses a 1 second wait period.
2. Two Inputs - In this method, multimodal input is assumed to consist of two inputs from two modalities. After inputs from two modalities have been received, the integration process is performed and the result sent to the DM. A window of 3 seconds is used after receiving the first input. (Oviatt et. al. 1997)
3. Information evaluation - In this method integration is performed after receiving each input, and the result is evaluated to determine if the information can be transformed to a command that the system can understand. If transformation is possible, the work of MMIF is deemed complete and the information is sent to the DM. In the case of an incomplete transformation, a windowing technique is used. This approach is similar to that of Vo and Waibel (1997).

4 Use case study

We used a multimodal in-car navigation system (Gupta et. al., 2002), developed using the MMIF module and a dialogue manager (Thompson and Bliss, 2000) to perform this study. Users can interact with a map-based display to get information on various locations and driving instructions. The interaction is performed using speech, handwriting, touch and gesture, either simultaneously or sequentially. The system was set-up on a 650MHz computer with 256MB of RAM and a touch screen.

Figure 1: Multimodal navigation system

4.1 Subjects and Task

The subjects for the study were both male and female in the age group of 25-35. All the subjects were working in technical fields and had daily interaction with computer-based systems at work. Before using the system, each of the subjects was briefed about the tasks they needed to perform and given a demonstration of using the system.

The tasks performed by the subjects were:

- Dialogue with the system to specify a few different destinations, e.g. a gas station, a hotel, an address, etc. and
- Issue commands to control the map display e.g. zoom to a certain area on the map.

Some of the tasks could be completed both unimodally or multimodally, while others required multiple inputs from the same modality, e.g. providing multiple destinations using touch. We asked the users to perform certain tasks in both unimodal and multimodal manner. The users were free to choose their preferred mode of interaction for a particular task. We observed users' behavior during the interaction. The subjects answered a few questions after every interaction on acceptability of the system response. If it was not acceptable, we asked for their preference.

4.2 Observations

The following observations were made during and after analysis of the user study based on aggregate results from using all the three methods of collecting multimodal input.

Multimodality

These observations were of critical importance to understand the nature of multimodal input.

- Multimodal commands and dialogue usually consisted of two or three segments of information from the modalities.
- Users tried to maintain synchronization between their inputs in multiple modalities by closely following cross-modal references with the referred object. Each user preferred either to speak first and then touch or vice versa almost consistently, implying a preferred interaction style.
- Sometimes it took a long time for some modalities to produce a semantic representation after capturing information (e.g. when there was a long spoken input or when used on lower end machines). The MMIF module did not collect all the inputs in that turn because it received some input after a long time interval from the previous input(s).

User preference

- Users became impatient when the system did not respond within a certain time period and so they tried to re-enter the input when the system state was not being displayed to them.
- During certain stages of interaction, the user could only interact with the system unimodally. In those cases they preferred that the system does not wait.

Performance of various schemes

The performance of the various methods to predict the completion of the user turn depended on the kind of activity the user was performing. A multimodal command is defined as multimodal input that can be translated to a system action without the need for dialogue, for example, zooming in a certain area of a map. On the other hand, multimodal dialogue involved multi-turn interaction in which the user guided the system (or was guided by the system) to provide information or to perform some action.

- When a multimodal command was issued, the user preferred the "information evaluation" and "two input" methods. This was because most of the multimodal commands were issued using two modalities. The "Windowing" method suffered from a delayed response from the system. The user got the impression that the system did not capture their input.
- During multimodal dialogue the performance of the "two input" method was poor as sometimes a multimodal turn has more than two inputs. Multimodal dialogue usually did not result in the evaluation of a complete command so the performance of the "information evaluation" technique was similar to that of "Windowing".

Efficiency

- If users acted unimodally, then it took them longer than the average time required to provide the same information in multimodal manner.

4.3 Measurements

Several statistical measures were extracted from the data collected during the user study.

Multimodality

The total number of user turns was 112. 83% of them had multimodal input. This shows an over-

whelming preference for multimodal interaction. This is compared to 86% recorded in (Oviatt et. al. 1997). 95% of the time users used only two modalities in a turn. Usually there were multiple inputs in the same modality. Of the multimodal turns, 75% had only two inputs, and the rest had more than 2 inputs. To provide multimodal input, speech and touch/gesture were used 80% of the time, handwriting and gesture were used 15% of the time and speech and handwriting were used 5% of the time.

Temporal analysis

During multimodal interaction, 45% of inputs overlapped each other in time, while the remaining 55% followed the previous after some delay. This reinforces earlier recordings of 42% simultaneous multimodal inputs (Oviatt et. al. 1997). The average time between the start of simultaneous inputs in two different modalities was 1.5 seconds. This also matches earlier observations of 1.4 seconds lag between the end of pen and start of speech (Oviatt et. al. 1997). The average duration of a multimodal turn was 2.5 seconds without including the time delay to determine the end of turn. The average delay to determine the end of user turn during multimodal interaction was 2.3 secs.

Efficiency

We observed that unimodal commands required 18% longer time to issue than multimodal commands, implying multimodal input is faster. For example, it is easier to point to a location on a map using touch than using speech to describe it. A long sentence also decreases the probability of recognition. This compares favorably with observations made in (Oviatt et. al., 1997) which recorded a 10% faster task performance for multimodal interaction.

Robustness

We labeled as errors the cases where the MMIF did not produce the expected result or when all the inputs were not collected. In 8% of the observed turns, users tried to repeat their input because of slow observed response from the system. In another 6% of observed turns, all the input from that turn was not collected properly. 4% was due to an input modality taking a long time to process user input (possibility due to resource shortfall) and the remaining 2% were due to the user taking a long time between multimodal inputs.

5 Analysis

Following an analysis of the above observations and measurements, we came to the following conclusions:

- Multimodal input is segmented with the user making a conscious effort to provide synchronization between inputs in multiple modalities. The synchronization technique applied is unique to every user. Multimodal input is likely to have a limited number of segments provided in different modalities.
- Processing time can be a key element for MMIF when deploying multimodal interactive systems on devices with limited resources.
- Knowledge of the availability of current modalities and the task at hand can improve the performance of MMIF. Based on the current task for which the user has provided input, different techniques should be applied to determine the end of user turn.
- Users need to be made aware of the status of the MMIF and the modes available to them. A uniform interface design methodology should be used, allowing the availability of all the modalities during all times.
- Timing between inputs in different modalities is critical to determine the exact relationship between the referent and the referred.

5.1 Temporal relationship

Based on the observations, a fine-grained classification of the temporal relationship between user inputs is proposed. Temporal relationship is defined to be the way in which the modalities are used during interaction. Figure 2 shows the various temporal relationships between feature structures that are received from the modalities. A, B, C, D, E, and F are all feature structures and their extent denotes the capture period. These relationships will allow for a better prediction of when and which modality is likely to be used next by the user.

- **Temporally subsumes –** A feature structure X temporally subsumes another feature structure Y if all time points of Y are contained in X. In the figure D temporally subsumes E.
- **Temporally Intersects –** A feature structure X temporally intersects another feature structure Y if there is at least one time point that is contained in both of them. However, the end point of X is not contained in Y and the start point of Y is not contained in X. In the figure B and C temporally intersect each other.
- **Temporally Disjoint –** A feature structure X is temporally disjoint from another feature structure Y if there are no time points in common between X and Y. In the figure, B and F are temporally disjoint.
- **Contiguous –** A feature structure X is contiguous with another feature structure Y if X starts immediately after Y ends. The two events have no time points in common, but there is no time point between them. For example, in the figure A is contiguous after B.

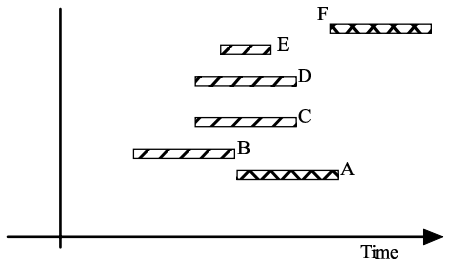

Figure 2: Feature structure temporal relationships

6 Enhancement to MMIF

It was proposed to augment the MMIF component with a wait mechanism that collects information from input modalities and adaptively determines the time when no further input is expected. The following factors were used during the design of the adaptive wait mechanism:

1. If the modality is specialized (i.e. it is usually used unimodally) then the likelihood of getting information in another modality is greatly reduced.
2. If the modality usually occurs in combination with other modalities then the likelihood of receiving information in another modality is increased.
3. If the number of segments of information within a turn is more than two or three then the likelihood of receiving further information from other modalities is reduced.
4. If the duration of information in a certain modality is greater than usual, it is likely that the user has provided most of the information in that modality in a unimodal manner.

6.1 Dynamic Time Windows

The enhanced method is the same as the information evaluation method except, that instead of the static time window, a dynamic time window based on current input and previous learning is used.

Time Window prediction

A statistical linear predictor was incorporated into the MMIF. This linear predictor provided a dynamic time window estimate of the time to wait for further information. The linear prediction (see figure 2) was based on statistical averages of the time required by a modality i to process information ($AvgDur_i$), the time between modalities i and j becoming active ($AvgTimeDiff_{ij}$), etc. The forward prediction coefficients (c_i and c_{ij}) were based on the predicted modalities to be used or active, the current modality used, and the temporal relationship between the predicted and current modality.

$$TTW = \sum_{i=1}^{n} c_i AvgDur_i + \sum_{i \neq j}^{n} c_{ij} AvgTimeDiff_{ij}$$

Figure 3: Linear prediction equation

Bayesian Learning

Machine learning techniques were employed to learn the preferred interaction style of each user. The preferred user interaction style included the most probable modality(s) to be used next and their temporal relationship. Since there is a lot of uncertainty in the knowledge of the preferred interaction style, a Bayesian network approach to learning was used. The nodes in the Bayesian network were the following:

a) Modality currently being used

b) Type of current input (i.e. type of semantic structure)

c) Number of inputs within the current turn

d) Time spent since beginning of current turn (this was made discrete in 4 segments)

e) Modality to be used next

f) Temporal relationship with the next modality

g) Time in current modality greater than average (true or false)

Learning was applied on the network using data collected during previous user testing. Learning was also applied online using data from previous user turns thus adapting to the current user.

7 Results

The enhanced module was tested using the data collected in previous tests and further online tests. The average delay in determining the end of turn reduced to 1.3 secs. This represents a 40% improvement on the earlier results. Also based on online experiments, with the same users and tasks, the number of times users repeated their input was reduced to 2% and collection errors reduced to 3% (compared to 8% and 6% respectively). The improvement was partly due to the reduced delay in the determination of the end of the user's turn and also due to prediction of the preferred interaction style. It was also observed that the performance increased by a further 5% by using online learning. The results demonstrate the effectiveness of the proposed approach to the robustness and temporal performance of MMIF.

8 Conclusion

An MMIF module with Dynamic Time Widows applied to an adaptive wait mechanism that can learn from user's interaction style improved the interactivity in a multimodal system. By predicting the end of a user turn, the proposed method increased the usability of the system by reducing errors and improving response time. Future work will focus on user adaptation and on the user interface to make best use of MMIF.

References

Anurag Gupta, Raymond Lee and Eric Choi. 2002. *Multi-modal Dialogues As Natural User Interface For Automobile Environment*. In Proceedings of Australian Speech Science and Technology Conference, Melbourne, Australia.

Anurag Gupta. 2003. *A Reference Model for Multimodal Input Interpretation.* In Proceedings of Conference on Human Factors in Computing Systems (CHI2003), Ft. Lauderale, FL.

Michael Johnston, Srinivas Bangalore, Gunaranjan Vasireddy, Amanda Stent, Patrick Ehlen, Marilyn Walker, Steve Whittaker, and Preetam Maloor. 2002. *MATCH: An Architecture for Multimodal Dialogue Systems.* In proceedings of 40th annual meeting of Association of Computational Linguistics (ACL-02), Philadelphia, pp. 376-383

Minh T. Vo and Alex Waibel. 1997. *Modelling and Interpreting Multimodal Inputs: A Semantic Integration Approach.* Carnegie Mellon University Technical Report CMU-CS-97-192. Pittsburgh, PA.

Robert Carpenter. 1992. *The logic of typed feature structures.* Cambridge University Press, Cambridge.

Sharon L. Oviatt, A. DeAngeli, and K. Kuhn. 1997. *Integration and synchronization of input modes during multimodal human-computer interaction.* In Proceedings of Conference on Human Factors in Computing Systems, CHI, ACM Press, NY, pp. 415–422.

Sharon L. Oviatt, Phil. R. Choen, Li Z. Wu, J. Vergo, L. Duncan, Bernard Shum, J. Bers, T. Holzman, Terry Winograd, J. Landay, J. Larson, D. Ferro. 2000. *Designing the user interface for multimodal speech and pen-based gesture applications: State of the art systems and future research directions.* Human Computer Interaction, 15(4), pp. 263-322.

Will Thompson and Harry Bliss. 2000. *A Declarative Framework for building Compositional Dialog Modules.* In Proceedings of International Conference of Speech and Language Processing, Beijing, China. pp. 640 – 643.