ISIS: A Multilingual Spoken Dialog System developed with CORBA and KQML agents

Helen Meng[1], Shuk Fong Chan[1], Yee Fong Wong[1], Tien Ying Fung[1], Wai Ching Tsui[1], Tin Hang Lo[1], Cheong Chat Chan[1], Ke Chen[3], Lan Wang[3], Ting Yao Wu[3], Xiaolong Li[3], Tan Lee[2], Wing Nin Choi,[2] Yiu Wing Wong,[2] P. C. Ching[2] and Huisheng Chi[3]

[1]Human-Computer Communications Laboratory,
[2]Digital Signal Processing Laboratory,
The Chinese University of Hong Kong,
Shatin, N.T., Hong Kong SAR, China
[3]National Key Laboratory for Machine Perception,
Peking University
Beijing, China
{hmmeng@se.cuhk.edu.hk}

Abstract

ISIS, which abbreviates Intelligent Speech for Information Systems, is a trilingual spoken dialog system (SDS) for the financial domain. It handles two dialects of Chinese (Cantonese and Putonghua), as well as English ¨ the predominant languages in our region. The system supports spoken language queries regarding stock market information and simulated personal portfolios. Real-time information is retrieved directly from a dedicated Reuters satellite feed. ISIS provides a system test-bed for our work in multilingual speech recognition and generation, speaker authentication, language understanding and dialog modeling. Furthermore, ISIS supports our initial explorations in: (i) CORBA's *interoperability* and *scalability* for SDS development; in conjunction with (ii) *asynchronous* human-computer interaction by delegation to KQML software agents.

1. Introduction

ISIS (Intelligent Speech for Information Systems) is a spoken dialog system (SDS) for retrieving real-time market information and managing simulated personal portfolios. As exemplified by many prominent systems developed previously, e.g. [Seneff et al., 1999; Rudnicky et al., 1999; Rosset et al. 1999], an SDS delineates a scope for research on spoken language technologies, ISIS provides a system test-bed for our research in multilingual speech recognition and generation, speaker authentication, language understanding and dialog modeling.

ISIS is a trilingual system, supporting English, Cantonese and Putonghua ¨ the predominant languages used in our region. The system resembles a virtual stockbroker, which can provide the user with real-time stock market information and personal portfolio information. It also handles simulated financial transactions. The financial domain is of particular interest to our region, which is one of the world¡¦s financil centers. Furthermore, the financial domain provides the necessary complexity for our research in spoken dialog systems. For example, number expressions abound and they need to be disambiguated and understood; new companies continue to be listed in the stock exchanges, which demands that our system be extensible to accommodate the new information.

This paper reports on our initial explorations in: (i) CORBA's *interoperability* and *scalability* for SDS development; in conjunction with (ii) *asynchronous* human-computer interaction by delegation to KQML software agents.

Previous work in dialog system architecture design include the GALAXY system [Seneff et al., 1999], which has a client-server architecture integrating with several information servers covering the weather, air travel and city navigation domains. GALAXY consists of a *hub* interacting with a number of servers via the hub script, which specifies the servers¡¦ hosts, ports, functionalities, etc. Another example is the Open Agent Architecture (OAA) described in [Cohen et al., 1994]. Here the user¡¦s request is executed in a distributed manner, which involves a *blackboard server* interacting with a community of cooperative software agents running on distributed machines.

2. CORBA

We have chosen to implement ISIS based on CORBA (Common Object Request Broker Architecture). CORBA is a suite of specifications produced by the Object Management Group (OMG), and aims to provide ease and flexibility for distributing components. Based on object orientation, CORBA¡¦s components can be shared among applications. This middleware also allows distributed object applications to *interoperate* across platforms through the network, by providing the IDL (Interface Definition Language) to communicate with different programming languages running on multiple operating systems. In addition, CORBA provides the ORB (Object Request Broker) which is a class library that handles communication between the IDL and the objects. ORB handles object location, request routing and result returning.

The ISIS implementation consists of a host of object-oriented server classes (speech recognition, language understanding, speech generation, etc.), that communicate with the client classes (the text I/O GUI and the speech I/O GUI) to handle user input over the Internet. The server classes run on multiple platforms ¨ the UNIX and Windows NT operating systems, and programming languages including Java, C and Visual C++. Hence the IDL helps achieve *interoperability*. By altering the IDL, new classes can be added and tested without affecting the existing architecture. This enhances the *scalability* of the architecture. Furthermore, the implementation does not require that all server classes be started for development. For example, without starting the speech recognizers, users can still access the ISIS system via the text input GUI. Additionally, by virtue of the ORB, and CORBA¡¦s IIOP (Internet and Inte-ORB Protocol), our browser-based client can communicate with the server classes over the Internet or local Intranets, while maintaining *location transparency* of the objects.

3. KQML

As mentioned previously, this work includes our initial exploration with asynchronous human-computer interaction by

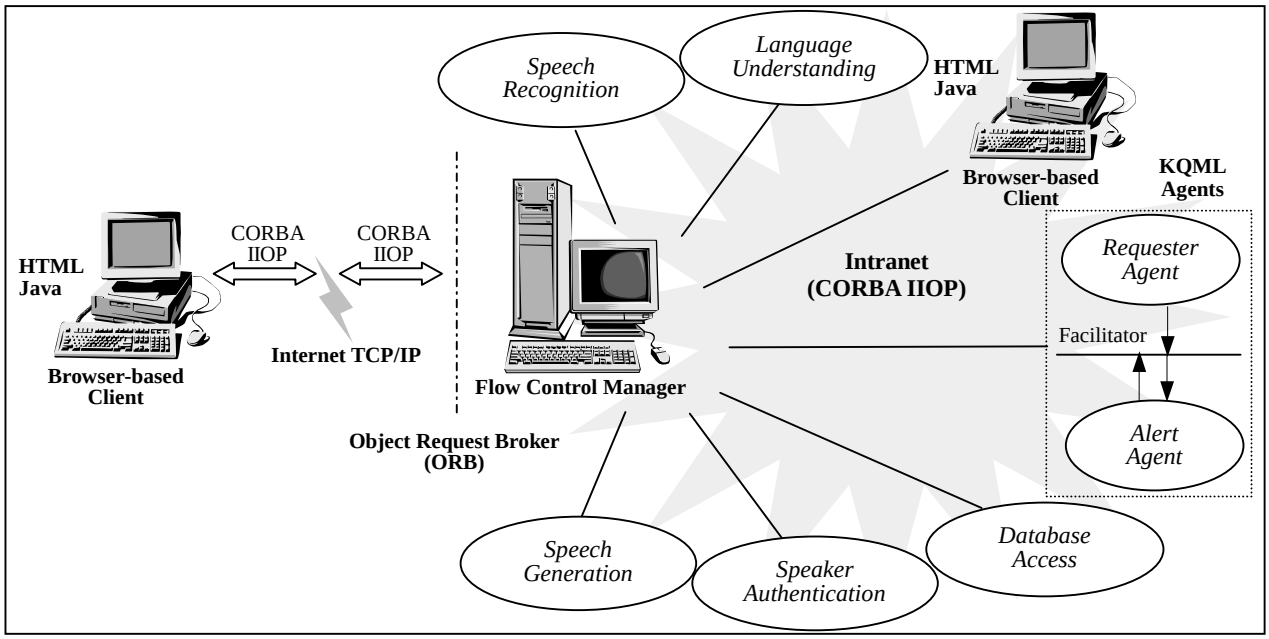

Figure 1: Overview of the ISIS System Architecture

delegation to KQML (Knowledge Query and Manipulation Language) software agents [Finin et al., 1994]. KQML is both a message format and a message-handling protocol to support information exchange and knowledge sharing among software agents. It enables a multi-agent communication with simple coding. KQML provides a basic architecture for us to implement the *user alert function* by means of information exchange among software agents. A non-blocking user query can result in a system reply (or an indeterminate number of replies) at irregular times in the future. In the mean time, the user can continue with the interaction session. We refer to this as *asynchronous* communication. Information exchange is mediated by a *facilitator,* which maintains a database of local services for routing messages.

4. System Architecture

Figure 1 illustrates the ISIS system architecture. There are six server classes H including the flow control manager, speech recognition, language understanding, database access, speech generation and speaker authentication. These communicate with a client class that supports text I/O and audio I/O by incorporating the applet with Java Sound API. These server and client classes inherit CORBA stubs and skeletons to communicate with each other by passing messages through the IDL. Additionally, there is a pair of KQML software agents that communicate via the facilitator to handle the user¡¦s price alert requests. When triggered, the Notifier agent issues a price alert request to the Monitor agent which tracks the market price continuously. When the user¡¦s specified price is met, the Monitor activates the Notifier to alert the user. The Notifier then sends the alert message to the Control Flow Manager.

Data is passed between the server/client classes in the format of XML (EXtensible Markup Language).[1] The data is labeled with descriptive semantic tags characterizing the server class operations. For example, Figure 2 shows the output of the natural language understanding (NLU) component based on the input *¡§buy five lots of HSBC at the market price please.`*

```
<NLU> <goal> buy </goal> <ric> 0005.HK </ric>
<num_lots> 5 </num_lots> <price> market </price> </NLU>
```

Figure 2: Example of a message produced by the language understanding server class. Input query *was ¡§buy five lots of HSBC at the market price please.o ric¡¨* denotes Reuters Instrument Code.

These XML tags are utilized by the Flow Control Manager for routing the data to the appropriate server/client classes for processing. The general flow control proceeds as follows: It begins with speech recognition, followed by natural language understanding, possibly speaker authentication (when the user is requesting personal portfolio information or conducting a financial transaction), information access (for real-time data captured from the Reuters satellite feed), and response generation (of text, speech and graphics). Additionally, the Flow Control Manager consists of seven Java-implemented helper classes:

(i) Semantic Frame Check H checks the NLU output for out-of-domain queries, and validates responses to confirmation sub-dialogs.

(ii) Context Manager Htakes care of discourse inheritance.

(iii) Data Check H checks that the necessary information attributes are provided before database access, as well as checks for valid data retrieved after database access.

(iv) Response Trigger H invokes text / spoken response generation when the appropriate data has been retrieved.

(v) Alert Manager H interfaces between the CORBA components and the KQML software agents. It passes alert requests and receives alerts from the agents.

(vi) System Log Manager Hmaintains a log of the entire dialog system flow.

(vii) Common Functions Manager H manages the functions of the Java-implemented classes mentioned above.

[1] http://www.w3.org/XML

5. Spoken Language Technologies

ISIS integrates a plethora of speech and language technologies, together with dialog modeling techniques. We will describe them in the following.

5.1 Speech Recognition

Speech recognition (SR) in ISIS aims to handle three languages (English, Cantonese and Putonghua (two dialects of Chinese). We have integrated an off-the-shelf English speech recognizer into our system. We have also developed preliminary versions of the Cantonese and Putonghua recognizers. These HMM recognizers use acoustic models based on the initials (I) and finals (F) of the dialects. The Chinese recognizers use a two-pass search, where the first pass creates a syllable lattice, and the second traverses the lattice with a language model to produce recognized word sequences.

5.2 Language Understanding

We applied our framework based on Belief Networks for natural language understanding [Meng et al., 1999]. Our approach involves parsing the user¡¦s query (which may be spoken or typed) by means of a semantic grammar. The semantic concepts obtained are then fed into a suite of Belief Networks (BN) for communicative goal inference. This approach was previously applied to the English sentences in the ATIS domain.

We have extended this natural language understanding (NLU) framework to handle Chinese as well as English queries, and ported from ATIS to the stocks domain. Details are reported in [Tsui & Meng, 2000]. Chinese input queries are first tokenized into a word sequence based on a 1100-word lexicon and a greedy algorithm. Parallel English and Chinese grammars were developed for semantic parsing, and they share a unified set of semantic concepts. We have also identified ten domain-specific goals for the stocks domain,[2] which include REAL-TIME QUOTES, NEWS, ORDER_AMENDMENTS, BUY, SELL, PORTFOLIO_INQUIRY, etc. Hence ten BNs were developed for goal inference. Should all ten BNs vote negative for a given query, it is rejected as out-of-domain (OOD).

The stocks domain has additional complexities for natural language understanding. Verbalized numbers abound in the domain-specific queries, and they can refer to stock codes, prices, number of lots, number of shares, etc. Consider the query example:

Buy HSBC, three hundred lots at a hundred and five per share.

Verbalized numbers are parsed to obtain their numeric values, and the number expressions are classified into the appropriate semantic category with considerations of *both* left and right contexts. For example, *a hundred and five*¡¨ above is a SHARE_PRICE. We find that this strategy is also useful for identifying the semantic category of out-of-vocabulary words, e.g. new stock names. Presently, our language understanding component only uses the *single* best recognition hypothesis is used for understanding. We plan to incorporate better SR/NLU coupling strategies as a next step.

5.3 Spoken Response Generation

System responses in ISIS need to be generated in three languages as well. For English, we generate the response text and sent it to the FESTIVAL system [Taylor et al, 1997], which has been integrated into ISIS. Language generation for English takes a response semantic frame as input. The frame specifies the language (i.e. English), the response type (e.g. STOCK_QUOTES_RESPONSE), related information attributes (e.g. STOCK_NAME, BID/ASK PRICES, etc.) and their values. Based on this response frame, the English response grammar is run in generative mode to produce the response text, to be used in text-to-speech synthesis.

For the Chinese dialects, we have designed and developed a meaning-to-speech generation methodology as described in [Fung & Meng, 2000]. The input Chinese response frame specifies the language (i.e. Cantonese or Putonghua) as well as the response type and related informational attributes. Based on this response frame, the Chinese response grammar is run in generative mode to specify the syllables (or syllable sequences) which need to be concatenated in sequential order. Our concatenative resequencing algorithm searches a bank of tonal syllable acoustic wave files, and selects units such that in a contiguous pair the right coarticulatory context of the left neighbor is compatible with the left coarticulatory context of the right neighbor. Coarticulatory context is specified in terms of distinctive features. The same approach is used for both Cantonese and Putonghua. We have demonstrated that this methodology can generate Chinese spoken responses that are highly natural and intelligible.

5.4 Speaker Authentication

During an interaction, the user may attempt to access general stock market information, or personalized portfolio information. Inferred goals such as BUY, SELL, PORTFOLIO_INQUIRY, from language understanding will trigger the *speaker authentication* component. The objective of speaker authentication (SA) is to automatically verify the speaker's claimed identity by his/her voice. Hence it provides a secured gateway for the user to access personal financial information and conduct financial transactions. The current SA component is a *text-dependent* speaker verification system, i.e. the system provides a digit string for the speaker to utter. Previous work indicates that the *Gaussian Mixture Model* (GMM) performs well for SA [Reynolds, 1992]. So a GMM-based SA with 16 mixture components is adopted for ISIS.

Another critical problem for speaker verification is the decision-making procedure. In ISIS, we adopt the hypothesis testing approach. We use the likelihood ratio between the claimed speaker°s model and a background model from other speakers. After comparing this likelihood ratio with a preset threshold, the binary decision of accept / reject is made.

5.5 Dialog and Discourse

Discourse inheritance is achieved by using an electronic form (E-form) model [Meng et al., 1996]. Semantic concepts from the current user¡¦s query take precedence over previous query (queries) for NLU. A mixed-initiative dialog model is used, and it can prompt for missing information, offer price alert services, invoke confirmation sub-dialogs, etc. The system also has a list of meta-commands e.g. HELP, GOOD-BYE, UNDO, etc., to allow the user to navigate freely in the dialog space. Table 1 provides an example of an ISIS dialog, aimed to illustrate some of the system¡¦s functionalities

6. Conclusions and Future Work

This paper describes the design and development of ISIS, a trilingual spoken dialog system for the financial domain. The system has been under development for nearly a year. We have reported on our preliminary exploration with a CORBA implementation and KQML software agents for user alert functions. Future work includes better coupling of speech

[2] This is based on data collection from appropriate subjects for task definition.

recognition outputs and natural language understanding, data collection using the prototype system, and performance evaluation based on the domain-specific data.

User	**User logs into ISIS**
ISIS	Welcome to ISIS. Please input your query.
User	**I would like the exchange rate of the Singapore dollar please**
ISIS	Sorry, I don¡ft understand your request. I can provide only information about stocks such as real-time quotes, news and charts. I can also provide information about a simulated personal portfolio, and handle buy and sell transactions. *(ISIS rejects an out-of-domain query about foreign exchange.)*
User	**I would like to buy three lots of Cheung Kong Holdings please**
ISIS	Please provide the bid price. *(Language understanding infers that the above is a BUY request, i.e. GOAL=BUY. Semantic Frame Check shows that BUY requires three attributes: a stock, the number of lots/shares and a bid price. The last one is missing and the system prompts for it.)*
User	我想先看一看每日的走勢圖 **Translation: Can you show me the daily chart first**
ISIS	0001 長實請睇 Translation: 0001 (a stock code) Cheung Kong Holdings Ltd. Please check (daily chart shown). *(Language understanding infers that GOAL=SHOW_CHART. Semantic Frame Check shows that this requires a stock name and a chart type. The former is missing from the query but is inherited from discourse. The daily chart is shown on the graphical user interface.)*
User	**Buy three lots at eight nine dollars please**
ISIS	Please say the following digits in English to verify your identify: 1 3 7 6 4 *(Language understanding infers that GOAL=BUY and the necessary attributes are provided. The Flow Control Manager sees that this GOAL requires speaker verification.)*
User	**One, three, seven, six, four**
ISIS	Please confirm the following action: buy three lots of Cheung Kong Holdings Limited at eight-nine dollars. Please say yes to confirm or no to reject to action. *(SV gave a positive outcome. The user was allowed to proceed with the BUY request. Requests such as buy or sell always enter into a confirmation subdialog.)*
User	**Yes**
ISIS	The current bid price for Cheung Kong Holdings is at ninety dollars and ten cents. Do you want us to notify you when the price drops to eighty nine dollars? Please say yes or no. *(The system checks the real-time bid price to implement the BUY transaction, but found that the requested bid price is too low. It offers to launch a software agent to monitor the stock price for the user.)*
User	**Yes**
ISIS	The Alert Agent is launched! *(Time stamp for the agent is May 30 2000 at 3:59 PM)*
User	**Show me the monthly chart of HSBC**
ISIS	The monthly chart of HSBC is shown below *(graphic shown)* There is one alert message for you. The stock price of Cheung Kong Holdings Limited is at eighty nine dollars presently. Please confirm your previous request on May 30, 2000 at 3:59PM. You would like to buy three lots of Cheung Kong Holdings Limited at your requested price of eighty nine dollars per share. Please say yes to confirm or no to reject the action. *(The Notifier agent came back with an alert message.)*
User	**Yes**
ISIS	The buy transaction is completed. Please input a new query.
User	**Bye**
ISIS	Goodbye and see you later. *(BYE is a meta-command which can terminate the session at any time.)*

Table 1: An example dialog from the ISIS system.

7. Acknowledgments

This research is supported by the Joint Center for Intelligence Engineering between Peking University and the Chinese University of Hong Kong. We thank Reuters Hong Kong for donating their satellite feed to support our research.

8. References

1. Cohen, P. R., A. J. Cheyer, M. Wang, and S. C. Baeg, "An open agent architecture," in AAAI Spring Symposium, pp. 1-8, March 1994.

2. Finin, T., R. Friitzson, D. McKay and R. McEntire, ¡§KQML¡è A Language and Protocol for Knowledge and Information Exchange,¡¨ Technical Report C-94-02, University of Maryland, UMBC

3. Fung T. Y. and H. Meng, "Concatenating Syllables for Response Generation in Domain-Specific Applications," Proceedings of ICASSP, 2000.

4. Meng, H., W. Lam and C. Wai, "To Believe is to Understand," Proceedings of Eurospeech, 1999.

5. Meng H., et al., ¡§WHEELS: A Conversational System on Electronic Automobile Classified, Proceedings of ICSLP, 1996.

6. Reynolds, D. A., ¡§A Gaussian mixture modeling approach to text-independent speaker identification,¡¨ PhD thesis, Georgia Institute of Technology, 1992.

7. Rosset, S., S. Bennacef and L Lamel, ¡§Design Strategies for Spoken Language Dialog Systems,¡¨ Proceedings of Eurospeech, 1999.

8. Rudnicky, A. et al., ¡§Creating Natural Dialogs in the Carnegie Mellon Communicator System,¡¨ Proceedings of Eurospeech, 1999.

9. Seneff, S., et al., "Organization, Communication and Control in the Galaxy-II Conversational System,¡¨ Proceedings of ICSLP, 1998.

10. Taylor, P. et al., "The architecture of the Festival speech synthesis system", in Proceedings of the 3rd ESCA Workshop on Speech Synthesis, pp.147-151, 1997.

11. Tsui, C. and H. Meng, ¡§Comprehension Across Application Domains and Languages,¡¨ Proceedings of ISCSLP, 2000