Intelligent Speech for Information Systems: Towards Biliteracy and Trilingualism

Helen M. Meng[1]*, Steven Lee*[2] *and Carmen Wai*[1]

[1]Human-Computer Communications Laboratory
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong SAR, China
Tel: +852 2609 8327
Email: {hmmeng, cmwai}@se.cuhk.edu.hk

[2]SpeechWorks International Ltd.
695 Atlantic Avenue
Boston, MA 02111
Tel: +1.617.428.4444
Email: sclee@speechworks.com

Intelligent Speech for Information Systems: Towards Biliteracy and Trilingualism

Helen M. Meng[1], *Steven Lee*[2] *and Carmen Wai*[1]

[1]Human-Computer Communications Laboratory
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong SAR, China
Tel: +852 2609 8327
Email: {hmmeng, cmwai}@se.cuhk.edu.hk

[2]SpeechWorks International Ltd.
695 Atlantic Avenue
Boston, MA 02111
Tel: +1.617.428.4444
Email: sclee@speechworks.com

ABSTRACT

This paper reports on our research and development effort in human-computer spoken language interfaces, capable of processing English and Chinese, including two dialects for Chinese (Cantonese and Putonghua). This is the language environment in Hong Kong, and in order to develop human-computer spoken language interfaces that can be used by almost *anybody* in the region, we strive to develop speech and language technologies capable of handling biliteracy and trilingualism. The context of use is in accessing real-time information in the foreign exchange domain. Users can call and converse with our system using both fixed line telephones and mobile phones. Both have high penetration in Hong Kong, and the latter offers mobile access to real-time financial information.

Keywords

Speech interfaces, spoken language systems, multilingual, conversational systems

1. INTRODUCTION

This paper reports on our research and development effort in human-computer spoken language interfaces, capable of processing English and Chinese, including two dialects for Chinese – (i) Putonghua, the official Chinese dialect, and (ii) Cantonese, a major dialect spoken by tens of millions of people in Hong Kong, Macau, South China and many overseas Chinese communities.[1] This is the language environment in Hong Kong, where the official documents are written in English as well as Chinese; and the populace speaks Cantonese, Putonghua and English. Hence in order to develop human-computer spoken language interfaces which can be used by almost *anybody* in Hong Kong, we strive to develop speech and language technologies capable of handling biliteracy and trilingualism. The context of use is in accessing real-time information in the financial domain. Our system supports not only telephone access via landline phones, but also mobile phones, thus offering mobile access to real-time information.

Spoken language systems have previously been developed to support mixed-initiative dialog interaction in a multitude of application domains, which characteristically have several task-specific user goals and constraints. Examples include air travel (Price 1990), railway information (den Os et al., 1999), restaurant guide (Jurafsky et al., 1994), ferry timetables (Carlson, 1994), weather (Zue et al., 1997), electronic automobile classifieds (Meng et al., 1996), electronic assistants (Jeanrenaud et al., 1999) and tourist information (Deviller and Bonneu-Maynard, 1999). The languages concerned include English and a number of European languages. A few systems have also been developed for Mandarin Chinese (Yang and Lee, 1998).

[1] Putonghua has approximately 1,400 distinct syllables and four lexical tones, and a light tone; Cantonese has approximately 1,800 distinct syllables and nine lexical tones (Wong et al., 1999).

In this work, we have chosen the foreign exchange (FOREX) domain, which is well-suited for Hong Kong. The region has one of the largest foreign exchange trading centers in the world. The scope of our system covers the thirty two globally traded currencies included in the Reuters data feed, and the global nature of this application is appropriate for the development of a multilingual application. As mentioned earlier, we have also chosen to support telephone access via landline and mobile phones. Penetration of the former is near saturation and the latter is over 50%. We also support the multiple mobile phone standards provided in Hong Kong – PCS, GSM and CDMA, etc. (OFTA). Our ultimate goal is to explore the research issues involved in the development of a multilingual conversational interface, with the objective of achieving usability from the perspectives of language (trilingual) and location (fixed-line and mobile telephone access). As an initial step, we developed a bilingual system for the foreign exchange domain. At this stage, we focused on the following issues:

- Bilinguality – we began with Cantonese and English, the two predominant languages used in the region.
- Affordance of the dialog design – we aim to support effective interaction of both novice and expert users.
- Evaluation – we conducted a series of user trials and usability surveys to evaluate the end-to-end system.

Discourse handling is planned as a next step. Presently, our system has some capability of intelligent dialog modeling – it can prompt for disambiguation if the user merely specifies a currency name which may refer to the currency in multiple countries, e.g. the Franc (Belgian Franc, French Franc and Swiss Franc); or the Krone (Danish Krone and Norwegian Krone).

2. CU FOREX: SYSTEM ARCHITECTURE

CU FOREX is the name of our system (Meng et al., 2000). It supports inquiries about foreign exchange, including the bid / ask exchange rates between two currencies, and deposit interest rates for a particular currency at various time durations (twenty four hours, one week, one month, two months… one year). Real-time financial information is retrieved from the Reuters satellite feed by a data capture process and stored in a relational database. Figure 1 illustrates the overall architecture of the system.

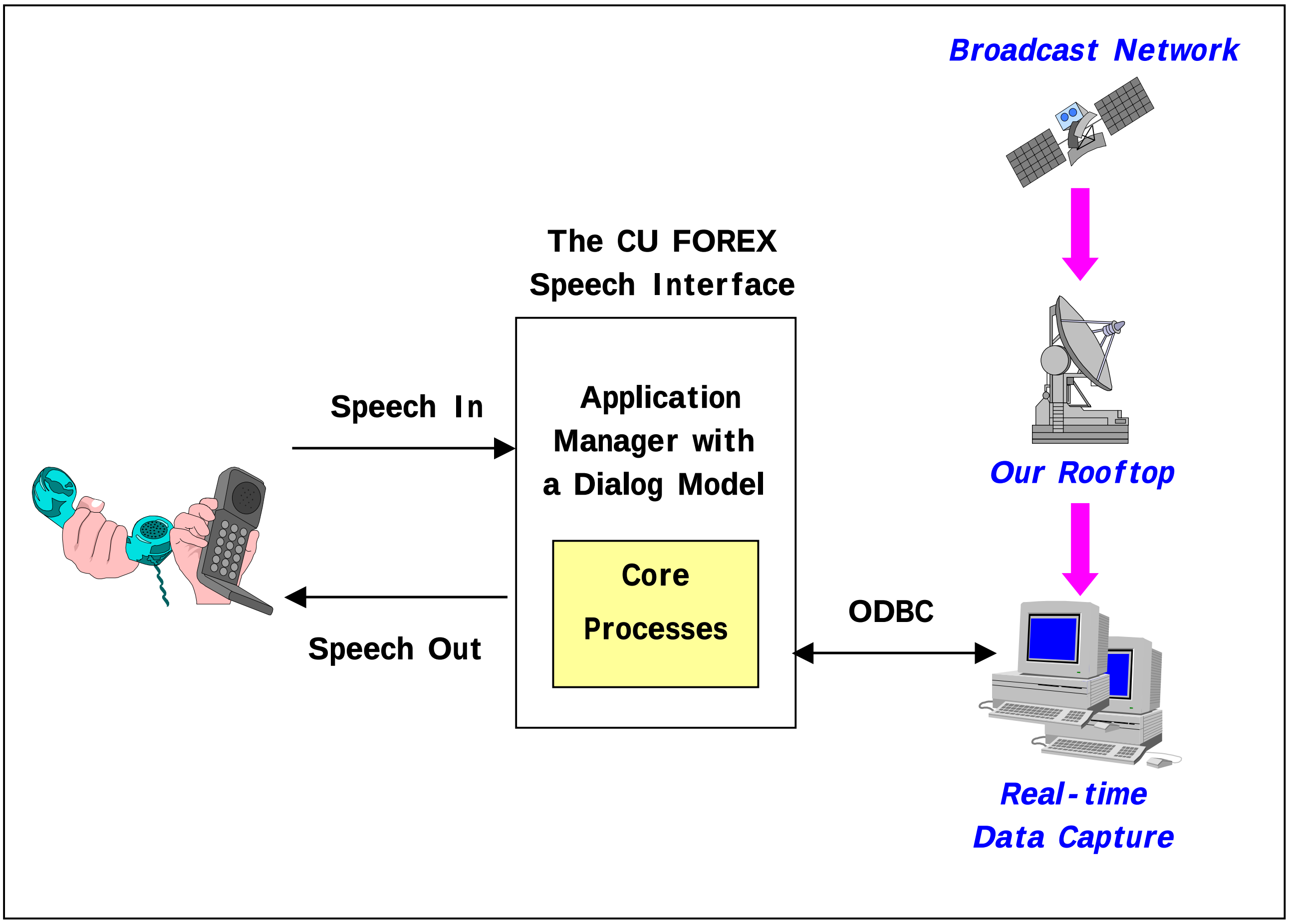

Figure 1: The CU FOREX system – overall architecture.

We receive a dedicate data feed from Reuters through a satellite dish mounted on the rooftop of our building. We have developed a software data capture component that continuously updates a relational database (SQL server) with the real-time data. The CU FOREX system communicates with the database via ODBC. Users can call up via a landline phone or mobile phone, and converse with the system to inquire about the foreign currency rates. In order to support such an interaction, the core processes include:

- bilingual speech recognition – current we handle Cantonese and English, but are extending to Putonghua.
- natural language processing – this processes user inputs in the form of complete questions.
- concatenative speech synthesis – this provides a spoken presentation of the raw data from Reuters in either Cantonese or English.

These are integrated with the application manager, which includes a dialog model. The interface is developed on a SpeechWorks 4.0 and InterVoice InVision platform, running on a Pentium II machine (300MHz) with 64M RAM.

2.1 Speech Recognition

Our speech recognition component handles both Cantonese and English. Our Cantonese transcription is based on the LSHK standard (LSHK), while our English transcription adopts the ARPABET phonetic labels. Our vocabulary has approximately 500 entries, covering country and currency names in the foreign exchange domain, as well as their (colloquial) variations, e.g. the "*German Mark*" and "*D-Mark*" both refer to the "*Deutsche Mark*". Similar examples include "馬仔" for "馬克", and the "*greenback*" for "*US Dollar*". To increase flexibility, users can call with either landline or mobile phones. The mobile phone service providers in Hong Kong adopt a number of standards, including PCS, GSM and CDMA. Hence our speech recognizer needs to handle these various types of data.

2.2 Natural Language Processing

We have developed parallel context-free grammars for natural language processing, one for English and the other for Chinese. This is necessary as English and Chinese have rather different word orders. As an illustration, consider the following query pairs:

English: *What is the exchange rate between the US dollar and the Hong Kong dollar?*

Chinese: 我想問美金兌港幣▫既匯率係幾多
Approximate Translation for the Chinese query:
<US DOLLAR><VERSUS><HK DOLLAR><EXCHANGE RATE>

Natural language is bypassed should the user select a directed dialog interaction. In a directed dialog interaction, the user inputs only one informational attribute per dialog turn.

2.3 Dialog Modeling

The system provides a directed dialog interaction for *novice* users, and natural language shortcut interaction for *expert* users. The dialog structure is consistent across the two languages. Example dialogs are shown in Tables 1 and 2. A directed dialog guides the user through a session, and elucidates what may be said at various stages of the interaction. User feedback is enriched by the use of audio icons to distinguish among the various menu choices.[2] Novice users may also take control and self-navigate around the dialog tree by the use of meta-commands, which include:

[2] The idea is contributed by Brian Lawrence from SpeechWorks.

HELP 求助: context-sensitive help instructions

REPEAT 重覆: repeat the last generated response

START AGAIN 重新開始: return to the main menu

CHANGE LANGUAGE 轉換語言: switch to the other language and start again from the main menu

GOODBYE 拜拜: end the session

The natural language interaction is tailored for expert users who may wish to expedite the inquiry session by uttering a full query, and traverse the entire session within one interaction.   Hence the query may carry multiple attributes for database retrieval, as opposed to a single attribute per utterance in the directed dialog case.

Both directed dialog and natural language interaction produce a semantic frame to invoke database access. The verbalized (spoken) response is generated by template-based word concatenation.

System	Welcome to CU FOREX Hotline. Which language would you prefer, “Cantonese” or “English”? 歡迎致電寰宇之聲.請講出你想用▫既語言 “廣東話” 定係 “英文”?
Caller	**English**
System	This system will guide you through a directed dialog to accomplish your enquiry. Main Menu. Would you like “exchange rates” or “interest rates”?
Caller	**Interest rate.**
System	Interest rate menu. Which currency are you interested in?
Caller	<UNRECOGNIZED>
System	Sorry, I didn’t understand you. Please say the currency name or the country name again.
Caller	**Australian Dollar.**
System	I think you said Australian Dollar. Is that correct? Please say “yes” or “no”.
Caller	**Yes.**
System	Would you like deposit durations for twenty four hours, one month…. or one year? Or, you can say “all of them” to hear all the interest rates quotes.
Caller	**One month.**
System	The quotes we provide are for reference only. Interest rate, Australian dollar for one month is four point nine. Main menu. Please select “exchange rates” or “interest rates”.
Caller	**Change language.**
System	請問你想查詢邊種服務, “外幣匯率” 定係 “存款利率”
Caller	**外幣匯率**
System	請問你想查詢▫既匯率是 “兌港幣” 還是 “兌其他外幣”?
Caller	**兌港幣**
System	請問你想查詢邊種貨幣▫既匯率?
Caller	**丹麥克郎**
System	以下資料只供參考. 丹麥克郎兌港幣,買入價係 1.087 ,賣出價係 1.086.

Table 1. Example of a Directed Dialog Interaction.

System	Welcome to CU FOREX Hotline. Which language would you prefer, “Cantonese” or “English”? 歡迎致電寰宇之聲.請講出你想用▫既語言 “廣東話” 定係 “英文”.
Caller	**English**
System	This system can handle a full question for your enquiry. Main Menu. What kind of currency information are you interested in?
Caller	**I’d like to know the exchange rate between the Greenback and the German Mark please.**
System	The quotes we provide are for reference only. Exchange rate. US Dollar to Deutsche Mark. The buying rate is two point one two. The selling rate is two point one one. What kind of currency information are you interested in?
Caller	**Change language.**
System	請講出你想查詢▫既外匯資料.
Caller	**澳幣 一星期▫既存款利率**
System	以下資料只供參考 澳幣一星期存款利率係 4.82

Table 2. Example Dialog of a Natural Language Interaction.

3. EVALUATION

CU FOREX is currently made available to the public for experimentation. We run two phone lines constantly throughout the day – one for the directed dialog interaction (+852.2603.7884) and the other for natural language shortcut interaction (+852.2603.7830). Calls are recorded at certain times and the data acquired is used for usability studies. We have recruited 89 subjects[3] over a three-week period to conduct an evaluation of the system. All our subjects were interacting with a spoken language system for the first time. They were asked to refer to the system’s homepage on the Web (CU FOREX)[4], to obtain some brief information about our system. Each evaluator was asked to formulate several queries related

[3] Our evaluators are students from the Chinese University of Hong Kong.

[4] http://www.se.cuhk.edu.hk/hccl/demos/cu_forex, also included in the citations list of this paper.

to foreign exchange prior to calling the system. Our analysis is based on system logs, as well as questionnaires returned by our evaluators. We received a total of 423 foreign exchange queries in all, with a breakdown tabulated in Table 3.

Directed Dialog Queries 277				**Natural Language Queries** 146			
Cantonese 112		**English** 165		**Cantonese** 56		**English** 90	
Ex. 76	**Int.** 36	**Ex.** 88	**Int.** 77	**Ex.** 33	**Int.** 23	**Ex.** 53	**Int.** 37

Table 3. Breakdown of queries from our evaluators. Ex. and Int. stands for Exchange Rate and Interest Rate queries respectively.

3.1 The KAPPA Statistic

Based, on this corpus, we adopted the PARADISE framework (Walker et al., 1997) for our evaluation. The PARADISE framework offers a way to evaluate task completion with considerations in task complexity. We organized our evaluation data into Attribute Value Matrices (AVMs), where the columns are reference values to the task attributes, and rows are hypothesized values to the task attributes. Our attributes include LANGUAGE, EXCHANGE_RATE, INTEREST_RATE, CURRENCY_TO_BUY, CURRENCY_TO_SELL, CURRENCY_FOR_DEPOSIT and TIME_DURATION, and their values include bilingual lexical items. For a given confusion matrix *M* with total count *T*, the kappa (κ) coefficient measures the rate of actual agreement between the reference and hypothesized values, *P(A)*, normalized by the rate of agreement by chance, as shown in Equation (1).

$$K = \frac{P(A) - P(E)}{1 - P(E)} \text{..........(1)}$$

where P(A) and *P(E)* are computed according to Equations (2) and (3), and t_i is the sum of counts in column *i* of the AVM.

$$P(A) = \frac{\sum_{i=1}^{n} M(i,i)}{T} ......(2) \qquad P(E) = \sum_{i=1}^{n} (\frac{t_i}{T})^2 ....(3)$$

3.3 Comparison between Interaction Styles

We compared between the interaction styles of directed dialog (DD) and natural language (NL). We expect that the recognition performance of DD is better, leading to a higher task completion performance and kappa value. We have also measured the average transaction time per interaction session. Results are shown in Table 4.

Task Completion	**Directed Dialog Interaction**	**Natural Language Interaction**
Kappa statistic (κ)	0.938	0.876
Average Transaction Time	2.15 min	1.94 min

Table 4. Comparison between two interaction styles in terms of success rate and transaction times for task completion.

From Table 4 we see that the kappa success rate for DD is higher than NL. Analysis shows that the inferior performance of NL is due to (i) the higher difficulty in recognizing full queries versus utterances each with a single keyword. In addition, system re-confirmation is absent in the NL interaction. (ii) parse failures of the recognized query – we will elaborate on this point later.

Task failures are mostly caused by recognition errors, out-of-domain queries (e.g. the “Finland Markka”) and parse failures from the NL interaction. Parse failures may be caused by disfluencies in the user's speech (e.g. false starts, filled pauses, repairs, etc.) Average transaction time for the DD interaction is 12 seconds longer than that of the NL interaction, which implies that NL expedites the transaction to some extent. The difference is smaller than expected, mostly due to the greater latency in recognizing a full query compared to a short, keyword-based utterance.

3.4 Comparison between Subtasks

We have also compared the subtasks of exchange rate and interest rate enquiries. Results are shown in Table 5. While the transaction times are comparable between the two subtasks, the kappa values for the

	Exchange Rate Enquiries	**Interest Rate Enquiries**
DD, kappa value	0.939	0.930
DD, duration	2.13 min	2.19 min
NL, kappa value	0.877	0.869
NL, duration	1.95 min	1.94 min

Table 5. Comparing the task completion success rates and durations between exchange rate and interest rate enquiries.

exchange rate enquiries are higher than those for interest rate enquiries. We believe this is due to a more confusable vocabulary (i.e. the lexical items pertaining to the specification of time durations, where nearly all the keywords take the format of `<digit> <hours | week | months>`).

3.5 Comparison across Languages

Comparison between the Cantonese query set and English query set yields the results as displayed in Table 6.

	Cantonese	English
DD, kappa value	0.950	0.926
DD, duration	2.04 min	2.24 min
NL, kappa value	0.812	0.914
NL, duration	1.68 min	2.07 min

Table 6. Comparison of success rates and durations for task completion between the use of Cantonese versus English.

The choice of language does not seem to affect task completion duration, and the kappa values based on the DD interactions are close between the two languages. However, the NL interaction using Cantonese is noticeably worse than that of English. Data for NL interaction in Cantonese is relatively sparse, but analysis shows that we received some input phrasal (colloquial) structures which were not anticipated during grammar writing, e.g.

請問日圓兌加元而家係幾多

一蚊港銀兌幾多

我想問澳元存款利率一年

日圓兌澳元,唔該

The situation should improve as we extend our grammar.

4. TOWARDS BILITERACY AND TRILINGUALITY

We are extending our system to Putonghua to become a trilingual system. It is observed that a multilingual system may be beneficial for this application context – users generally know the set of the globally traded currencies, but not all the country/currency names in a single language. Hence multilinguality offers enhanced flexibility to support the user's inquiries.

Additionally, we aim to scale up our system to application domains of higher complexities, e.g. securities domain and financial news. Figure 2 illustrates the architecture of such as system. It is enhanced with a speaker verification component (Meng, 2000) (Meng et al., 2000) to secure access to private or personal financial information. The speech recognition and speech generation components are *trilingual*, while the language understanding component is *biliteral* for handling English / Chinese text coming from the recognizers. The remaining components in the system remain *language independent*.

4.1 Biliteral Language Understanding

We have devised a methodology that can semi-automatically induce a context-free grammar from un-annotated text corpora (Meng and Siu, 2000). The methodology has been demonstrated to work for both English and Chinese textual queries. Grammar induction is an agglomerative word clustering approach which can capture semantic categories as well as phrasal structures. The induction algorithm is amenable to prior human knowledge injection to catalyze the induction process. Moreover, the induced grammars are amenable to hand refinement as a post-process. The induced grammars can then couple with a parser to analyze natural language input and extract meaning from the user's query. The semi-automatic nature of the approach enhances portability across domains and languages. Our experiments have shown encouraging results when the induced grammar is compared with a handcrafted grammar for understanding.

4.3 Trilingual Speech Recognition and Generation

We are integrating three monolingual recognizers to effect trilingual speech recognition. As regards speech generation, the objective is to generate a spoken presentation of the raw data (numbers and codes) for the user, and maximizing the degrees of intelligibility and naturalness in the output acoustics. We have integrated the FESTIVAL (Taylor et al., 1998) speech synthesizer to synthesize English responses. However, since Chinese speech synthesizers are not easily available, we have developed our own corpus-based concatenative synthesis system, and applied the same technique to *both* Cantonese and Putonghua (Fung and Meng, 2000). We have chosen the syllable as our basic unit for synthesis, since the Chinese language is monosyllabic in nature. Each syllable unit is appended with two digits to encode the distinctive features in its left and right co-articulatory context. Our concatenative synthesis technique aims to maximize the intelligibility and naturalness of the generated acoustics within the scope of the domain. A listening test based on 12 subjects showed that the concatenative approach compares favorably with a domain-independent PSOLA synthesizer based on intelligibility and naturalness.

5. Conclusions and Future Work

In this paper, we have reported on the design, development and evaluation of the first version of our bilingual spoken dialog system in the foreign exchange domain – CU FOREX. System performance is measured based on the kappa measure and time durations for task completions, and the effects of interaction styles (directed dialog versus natural language), choice of language, and difference in subtasks were all considered. Continual developments include improving the speech recognition performance based on the data collected from real users, and refining our grammar for parsing spoken language with colloquialisms and disfluencies. We have also completed a preliminary end-to-end Putonghua version of the application, to be integrated with the main system to achieve trilinguality. Ongoing work includes semi-automatic grammar induction for natural language processing, and concatenative speech synthesis which is applicable across Chinese dialects to generate highly natural and intelligible spoken responses from the system. We will also research into strategies for dialog designs, which should be scalable in the comprehensibility, predicability and controllability of the interface, to fit the needs of novice, knowledgeable and expert users.

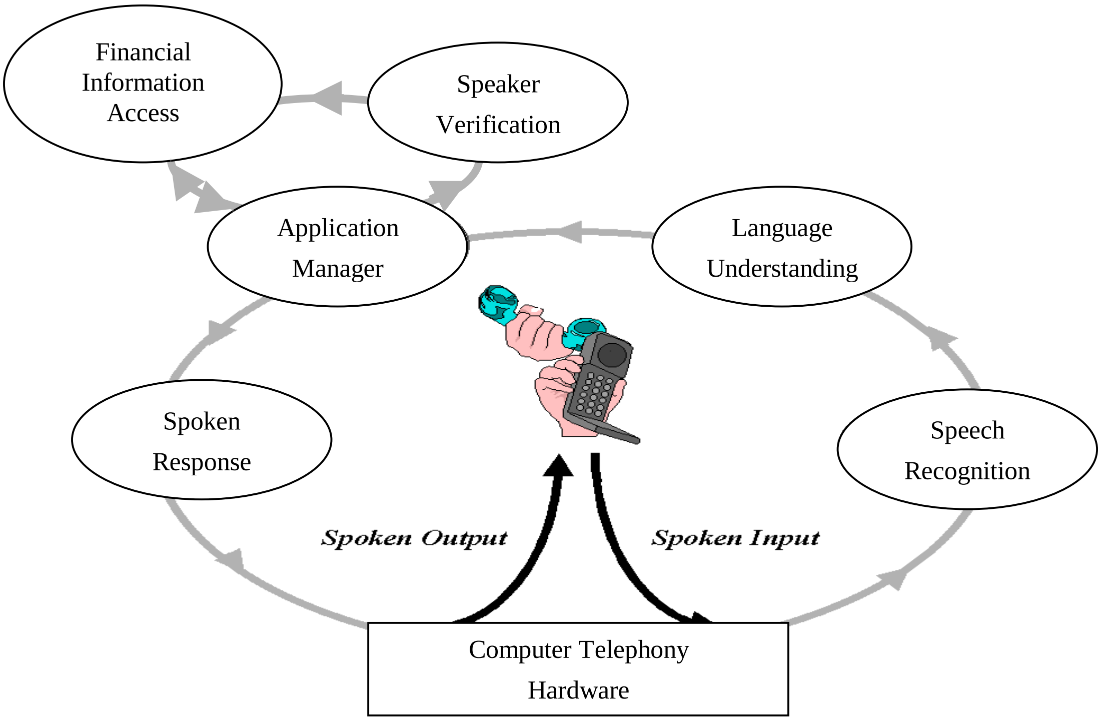

Figure 2. System Architecture of a Spoken Language Interface for Financial Information Access.

ACKNOWLEDGMENTS

We acknowledge our collaborators, SpeechWorks International Ltd. and IVRS (International) Ltd., on the development of CU FOREX system. We thank Reuters Hong Kong for donating the real-time data feed via satellite. The work reported here involves projects conducted by various members in the Human-Computer Communications Laboratory, Chinese University of Hong Kong. The speaker verification work is conducted in collaboration with the National Key Laboratory for Machine Perception, Peking University.

REFERENCES

1. Carlson, R.; (1994); "Recent Developments in the Experimental WAXHOLM Dialog System"; Proceedings of the ARPA Human Language Technology Workshop, pp.207-212; Morgan Kaufman.

2. CU FOREX, http://www.se.cuhk.edu.hk/hccl/demos/cu_forex/.

3. den Os, E., Boves, L., Lamel, L., Baggia, P., (1999); "Overview of the Arise Project Proceedings of 6th European Conference on Speech Communication and Technology; (cd-rom).

4. Deviller, L. and Bonneau-Maynard, H.; (1999); "Evaluation of Dialog Strategies for a Tourist Information Retrieval System"; Proceedings of 6th European Conference on Speech Communication and Technology; (cd-rom).

5. Fung, T. Y. and Meng, H.; (2000); "Concatenating Syllables for Response Generation in Spoken Language Applications"; Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing; (cd-rom).

6. Jeanrenaud, P., Cockroft, G., VanderHeidjen, A.; (1999); “A Multimodal, Multilingual Telephone Application: The Wildfire Electronic Assistant”; Proceedings of 6th European Conference on Speech Communication and Technology; (cd-rom).

7. Jurafsky, D., Wooters, C., Tajchman, G., Segal, U., Stolcke, A., Fosler, E., and Morgan, N.; (1994); “The Berkeley Restaurant Project,” Proceedings of the International Conference on Spoken Language Processing; (http://www.icsi.berkeley.edu/real/berp.html).

8. Lingustic Society of Hong Kong; (1997); Hong Kong Jyut Ping Character Table, Linguistic Society of Hong Kong Press.

9. Meng, H.; (2000); “Initial Development Towards a Trilingual Speech Interface for Financial Information Inquiries”; International Journal on Speech Technology; Volume 3, Issue 2, pp. 83-91.

10. Meng, H., Chan, S. F., Wong, Y. F., Fung, T. Y., Tsui, W. C., Lo, T. H., Chan, C. C., Chen, K., Wang, L., Wu, T. Y., Li. X., Lee, T., Choi, W. N., Wong, Y. W., Ching, P. C., and Chi, H.S.; (2000); "ISIS: A Multilingual Spoken Dialog System developed with CORBA and KQML agents”; Proceedings of the International Conference on Spoken Language Processing; (cd-rom).

11. Meng, H., Lee, S. and Wai, C.; (2000); “CU FOREX: A Bilingual Spoken Dialog System for Foreign Exchange Inquiries”; Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing; (cdrom).

12. Meng, H., and Siu, K. C.; (2001); “Semi-Automatic Acquisition of Domain-Specific Semantic Structures,” IEEE Transactions on Knowledge and Data Engineering, in press.

13. OFTA, Office of the Telecommunications Authority, Hong Kong SAR Government, http://www.ofta.gov.hk.

14. Price, P.; (1990); “Evaluation of Spoken Language Systems: the ATIS Domain,” Proceedings of the DARPA Speech and Natural Language Workshop; pp.91-95; Morgan Kaufman.

15. Taylor, P., Black, A., and Caley, R.; (1998); “The Architecture of the Festival Speech Synthesis System”; Proceedings of the 3rd ESCA/ COCOSDA Workshop on Speech Synthesis; pp. 147-151.

16. Walker, M., Litman, D., Kamm, C. and Abella, A.; (1997); “PARADISE: A Framework for Evaluating Spoken Dialogue Agents”; Proceedings of the 35th Annual Meeting of the Association of Computational Linguistics; (http://www.research.att.com/~walker/).

17. Yang Y., and Lee, L. S.; (1998); “A Syllable-based Chinese Spoken Dialogue System for Telephone Directory Services Primarily Trained with a Corpus,” Proceedings of the International Conference on Spoken Language Processing; (cd-rom).

18. Zue, V., Seneff, S., Glass, J., Hetherington, L., Hurley, E., Meng, H., Pao, C., Polifroni, J., Schloming, R., and Schmid, P.; (1997); “From Interface to Content: Translingual Access and Delivery of On-Line Information”; Proceedings of the 5th European Conference on Speech Communication and Technology; pp. 2227-2230; ESCA.