Spoken Language System Development for the MASK Kiosk

J.L. Gauvain, S. Bennacef, L. Devillers, L. Lamel
LIMSI-CNRS, BP 133
91403 Orsay cedex, FRANCE
{gauvain,bennacef,devil,lamel}@limsi.fr

INTRODUCTION

Spoken language systems aim to provide a natural interface between humans and computers through the use of simple and natural dialogues, so that the users can access stored information. In this paper we present our recent activities in developing such a system for the ESPRIT MASK (Multimodal-Multimedia Automated Service Kiosk) project. The goal of the MASK project is to develop a multimodal, multimedia service kiosk to be located in train stations, so as to improve the user-friendliness of the service. The service kiosk will allow the user to speak to the system, as well as to use a touch screen and keypad. The role of LIMSI in the project is to develop the spoken language system which will be incorporated in the kiosk. High quality speech synthesis, graphics, animation and video will used to provide feedback to the user. We have developed a complete spoken language data collection system for this task. In the actual service kiosk the spoken language system has to be modified so that the multimodal level transaction manager can control the overall dialog. The main information provided by the system is access to rail travel information such as timetables, tickets and reservations, as well as services offered on the trains, and fare-related restrictions and supplements. Other important travel information such as up-to-date departure and arrival time and track information will also be provided. Eventual extensions to the system will enable the user to obtain additional information about the train station and local tourist information, such as restaurants, hotels, and attractions in the surrounding area.

SYSTEM DESCRIPTION

The main components of the spoken language system are the speech recognizer, the natural language component which includes a semantic analyzer and a dialog manager, and an information retrieval component that includes database access and response generation. While our goal is to develop underlying technology that is speaker, task and language independent, any spoken language system will necessarily have some dependence of the chosen task and on the languages known to the system. The spoken query is decoded by a speaker independent, continuous speech recognizer[4]. which makes use of continuous density HMM with Gaussian mixture for acoustic modeling and n-gram backoff language models. The recognition lexicon has on the order of 1000 words, containing 35 station names. To meet the MASK project goals the number of stations needs to be increased to 500. The n-gram statistics are estimated on the texts of spoken queries. Since the amount of language model training data is small, some grammatical classes (such as cities, days, months, etc) are used to provide more robust estimates of the n-gram probabilities. The output of the recognizer is passed to the natural language component. In our current implementation the output of the speech recognizer is the best word sequence, however, the recognizer is also able to provide a word lattice. The semantic analyzer carries out a caseframe analysis to determine the meaning of the query[1, 2], and builds an appropriate semantic frame representation. In this analysis, keywords are used to select an appropriate case structure for the sentence without attempting to carry out a complete syntactic analysis. The major work in developing the understanding component is defining the concepts that are meaningful for the task. The concepts for the MASK task are similar to those used in our L'ATIS system[6]. These concepts are train-time, fare, change, type, reserve, and service and have been determined by analysis of queries taken from the training corpora to augment our *a priori* task knowledge. The dialog history is used to complete missing information in the semantic frame and the dialog context may be used to provide default values for required slots. The response generator uses the semantic frame to generate a database request to the database management system, and presents the result of the database query and an accompanying natural language response to the user. A vocal response is optionally provided, along with the written and tabular information.

DATA COLLECTION

The collection of spoken language corpora is an important research area and represents a significant por-

Month	*Jan*	*Feb*	*Mar*	*Apr*	*May*
#speakers	12	42	78	106	113
#queries	208	1603	3825	6219	6853
total #words	1.6k	12.2k	29.0k	44.4k	48.5k
#distinct words	271	691	902	1015	1049
#new words	-	420	211	113	34

Table 1: MASK data collection status

tion of the work in developing a spoken language system. The use of additional acoustic and language model training data has been shown to almost systematically improve performance in continuous speech recognition[5]. Similarly, progress in spoken language understanding is closely linked to the availability of spoken language corpora. For MASK we have recorded 113 speakers for a total of 6853 queries. The cumulative number of subjects and queries recorded are shown in Table 1. The average sentence length is 8 words. Each query is transcribed and classified as "answerable without context" (13%) , "answerable given context" (67%), "politeness forms" (<1%) , "out of domain" (<1%), and "temporarily out of domain" (19%). This latter category refers to queries which were not treated in the version of the system used to collect the data, but will be treated in future versions.

EXPERIMENTAL RESULTS

The MASK spoken language system has been evaluated on 205 queries from 10 speakers recorded using the data collection system. The speech recognition word accuracy is 85%. Natural language understanding of the exact transcriptions of the same set of spoken queries, without removing spontaneous speech effects such as hesitations or repetitions, is 93%. The complete spoken language system has an understanding rate of 79%. We expect that as more data is collected the understanding rate will improve, as we observed for our L'ATIS system[6]. With a speech recognition word accuracy of 94% we observe essentially no degradation in performance using the recognizer output instead of typed exact transcriptions for the L'ATIS system.

A frequent understanding error is due to sentences that include 2 queries such as *"Je voudrais réserver, remontrez-moi les tarifs (I would like to make a reservation, show me the fares again.)"*. While we instantiate correctly the 2 caseframes, we are not yet able to treat this at the dialog level. Another common error arises when the user makes an implicit reference to a previous response given by the system. For example, the user may ask for an earlier departure time "*Je veux partir plus tôt*", without ever having specified a departure time. To treat this, we need to interpret the previous response(s) given by the system. We are currently working on improving the maintenance of the dialog history so as to be able to relax previously specified constraints, so as to be able to handle requests such as *"tous les trains" (all the trains)*.

SUMMARY

We have presented our recent activities in developing spoken language systems for the MASK kiosk. The MASK kiosk will be evaluated in the Gare St. Lazare in Paris, early in 1996. We also continue our work on two other systems in the travel domain, L'ATIS and RAILTEL. These prototype systems are used for data collection. Significant performance improvements have been obtained by collecting additional data which have been used to improve the acoustic and language models of the recognition component. Similarly, analysis of the understanding errors on new data enables us to incrementally improve the understanding component. Our experience with data collection is that as the system performance is improved, subjects tend to speak more naturally, enabling us to record more representative spontaneous speech.

References

[1] H. Bonneau-Maynard, J.L. Gauvain, D. Goodine, L.F. Lamel, J. Polifroni, S. Seneff, "A French Version of the MIT-ATIS System: Portability Issues," *Eurospeech'93.*

[2] S.K. Bennacef, H. Bonneau-Maynard, J.L. Gauvain, L. Lamel, W. Minker, " A Spoken Language System For Information Retrieval," *ICSLP'94.*

[3] S.K. Bennacef, F. Néel, H. Bonneau-Maynard, "An Oral Dialogue Model based on Speech Acts Categorization," *ESCA Workshop on Spoken Dialog Systems*, Vigsø, Denmark, Spring 1995.

[4] J.L. Gauvain, L.F. Lamel, G. Adda, M. Adda-Decker "Speaker-Independent Continuous Speech Dictation," *Speech Communication*, **15**, pp. 21-37, Sept. 1994.

[5] J.L. Gauvain, L.F. Lamel, G. Adda, M. Adda-Decker "The LIMSI Continuous Speech Dictation System: Evaluation on the ARPA WSJ Task," *ICASSP'94.*

[6] L. Lamel, S. Bennacef, H. Bonneau-Maynard, S. Rosset, J.L. Gauvain, "Recent Developments in Spoken Language Sytems for Information Retrieval," *ESCA Workshop on Spoken Dialog Systems*, Vigsø, Denmark, Spring 1995.

[7] L. Lamel, S. Rosset, S. Bennacef, H. Bonneau-Maynard, L. Devillers, J.L. Gauvain, "Development of Spoken Language Corpora for Travel Information", *Eurospeech'95.*