Recent Developments in Spoken Language Sytems for Information Retrieval

L.F. Lamel, S.K. Bennacef, H. Bonneau-Maynard, S. Rosset, J.L. Gauvain
LIMSI-CNRS, BP 133
91403 Orsay cedex, FRANCE
{lamel,bennacef,hbm,rosset,gauvain}@limsi.fr

ABSTRACT

In this paper we present our recent activities in developing systems for vocal access to a database information retrieval system, for three applications in a travel domain: L'ATIS, MASK and RAILTEL. The aim of this work is to use spoken language to provide a user-friendly interface with the computer. The spoken language system integrates a speech recognizer (based on HMM with statistical language models), a natural language understanding component (based on a caseframe semantic analyzer and including a dialog manager) and an information retrieval and response generation component. We present the development status of prototype spoken language systems for L'ATIS and MASK being used for data collection.

1. INTRODUCTION

Spoken language systems aim to provide a natural interface between humans and computers through the use of simple and natural dialogues, so that users can have easier access to stored information. In this paper we present our recent activities in developing systems for vocal access to a database information retrieval system, and focus on issues in portability across tasks and languages. Our initial spoken language system was a French version[2] of ATIS (Air Travel Information Service) a designated common task for data collection and evaluation within the ARPA Speech and Natural Language Program.[7] ATIS allows users to acquire information derived from the Official Airline Guide about fares and flights available between a restricted set of cities within the United States. The user can also ask about other related information, such as the meals served on the flight or the type of aircraft, and fare-related restrictions.

This system has been ported to another related application, in the context of the ESPRIT project MASK. 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 are developing 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 will be 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. We also plan to provide up-to-date departure and arrival time and track information. Eventual extensions to the system will enable the to user to obtain additional information about the train station and local tourist information.

A related application being developed for the LE MLAP project RAILTEL will provide train timetable and scheduling information over the telephone. In this case the caller will only be able to interact with the system by voice implying that changes in the user-interface will be needed to limit the information and guidance provided to the user by speech synthesis.

2. SYSTEM OVERVIEW

An overview of the spoken language system for information retrieval is shown in Figure 1. The main components 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, whose output is then 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

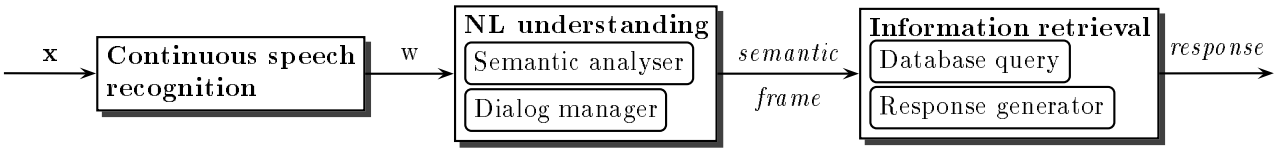

Fig. 1. Overview of the spoken language information retrieval system. **x** is the input speech signal, **w** is the word sequence output by the speech recognizer.

semantic analyzer carries out a caseframe analysis to determine the meaning of the query, and builds an appropriate semantic frame representation. 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 (DBMS), and presents the result of the database query and an accompanying natural language response to the user. A vocal response may be optionally provided for L'ATIS and MASK, and is the only mechanism for response in the RAILTEL system.

3. SPEECH RECOGNIZER

The speech recognizer is a large vocabulary, speaker independent, continuous speech recognizer[3]. Acoustic modeling makes use of continuous density HMM with Gaussian mixture of context-dependent phone models. For language modeling n-gram statistics are estimated on the transcriptions of the spoken queries. The decoder uses a time-synchronous graph-search strategy for a first pass with a bigram back-off language model (LM)[5]. The backoff component of the bigram is efficiently represented with a lexicon tree reducing the number of interword connections. The word recognition graph is built by putting together word models according to the grammar in one large HMM, where each word model is obtained by concatenation of phone models, according to its phone transcription in the lexicon. A trigram LM is used in a second acoustic decoding pass which incorporates the word graph generated in the first pass.

The recognition vocabulary for L'ATIS and MASK contain 826 words and 616 words, respectively. The initial MASK vocabulary was a modified version of the L'ATIS vocabulary where city/airport names were replaced with city/station names, and air travel related words replace with appropriate train travel domain words. The bigram and/or trigram language models are estimated on the texts of typed and spoken queries for each task. For L'ATIS, 6300 queries are used for LM training, allowing a class trigram to be estimated. Word classes are used for lexical items such as the cities, days, months, when there is no reason to believe that differences in their frequencies in the training data are significant or representative. Since only 2500 MASK queries were available for estimating the LM, only a class bigram is used.

4. UNDERSTANDING COMPONENT

The semantic analyzer carries out a caseframe analysis to determine the meaning of the query, and builds one or several semantic frames. This approach was described in detail for the L'ATIS task in [1]. One advantage of this approach is that the understanding procedure does not require verifying the correct syntactic structure of the sentence, but rather extracts the meaning using syntax as a constraint.

The major job in developing the understanding component is defining the concepts that are meaningful for the task. The three tasks dealt with in this work are all in a travel-related domain, and therefore share many commonalities. We have analyzed queries taken from the training corpora to augment our a priori task knowledge. The concepts for L'ATIS are flight-time, fare, stop, type, reserve. Most of these concepts are also found in the train travel domain, albeit with slightly different significations. For example, the concept type corresponding to aircraft type in ATIS corresponds to the type of train in MASK (TGV, TEE, etc), whereas the concepts related to arrival and departure times, and fares can be mapped directly. The MASK concepts are train-time, fare, stop, type, station, reserve, reductions and services.

A set of cases are used to represent the different types of information in all the caseframes. These cases may be classed according to the types of information: designation, locality, date, hour, fare and services. Casemarkers provide syntactic constraints necessary to extract the meaning of the request. These include pre- and post markers which are surface indicators designating a case. For example, in "**de** Paris", the preposition **de** designates Paris to be the departure city and in the phrase "à 14 **heures**", **heures** is an example of a postmarker, designating 14 to be a time. Pre- and post- casemarkers are not necessarily located next to the case, as in "qui **arrive** vers 22 heures", where 22 corresponds to the case arrival time, because it is preceded, although not directly, by the marker **arrive**.

An example of the caseframe train-time is given in Figure 2, where the KEYWORDS specify the words to select the given caseframe during parsing. The

CASEFRAME train-time {KEYWORDS: train, voyager, aller, partir... from: (quitte, de...) @city to: (a, pour, vers...) @city stop: (changement-a, via...) @city relative-departure-time: (partir+) avant, apres departure-time: (partir+) @hour-minute ...
CASEFRAME @city { city: paris, lille, lyon, marseille, nancy...}
CASEFRAME @hour-minute{...}

Fig. 2. Example caseframe.

Je veux aller demain matin de Paris à Marseille en passant par Lyon. (I would like to go from Paris to Marseille via Lyon tomorrow morning.)
<train-time> from: paris to: marseille stop: lyon relative-day: demain *(tomorrow)* morning-afternoon: matin *(morning)*

Fig. 3. Example semantic frame.

structure @city is the sub-caseframe and the case city contains a list of towns. hour-minute is also a sub-caseframe, but since understanding of numbers is very relevant to the travel information tasks (appearing also in dates and train numbers), a restricted local grammar is used to extract the corresponding values.

Figure 3 shows the resulting semantic frame for an example utterance. The keyword *aller* triggers the caseframe train-time, and the parser constructs the complete semantic frame by instanciating the slots *from, to* and *stop* with the corresponding words *Paris, Marseille* and *Lyon* respectively. The analysis is driven by the order in which the cases appear in the caseframe train-time.

5. RESPONSE GENERATION

Natural language responses are automatically generated from the semantic frame. Our initial approach returned essentially a paraphrase of the user's queries used to construct the semantic frame. This simplistic approach has the problem that the user may receive too much information, and find it difficult to extract the important part. We are currently modifying the response generation to provide more precise information to the user, in a simpler format. To do so, we have been experimenting with different forms of response - text strings, tables, and ticket images. Response generation and synthesis is of particular importance for RailTel where visual presentation of the information is not possible.

When vocal feedback is provided the speech has to be very natural and intelligible, as the average user cannot be expected to have previously heard synthetic speech, nor to be tolerant of poor quality output. In order to present variable information, we are using a speech concatenation approach[6] where the automatically generated response text is used to locate dictionary units for concatenation. This will be completed with a diphone dictionary constructed with speech from the same talker, so that in the event that the necessary dictionary units are not located, diphone synthesis can serve as a back-off mechanism. This capability can also enable the extension to new words. Simple playback of pre-recorded speech is used for fixed messages that are unlikely to be changed.

Month	*<Aug94*	*Sep-Oct*	*Nov-Dec*	*Jan- Feb*
#speakers	20	80	117	163
#queries	1111	3798	5625	8611
total #words	12.1k	39.7k	58.2k	86.7k
#dist. words	560	981	1132	1600
#new words	-	423	151	468

TABLE I. Progress in data collection for l'Atis.

6. DATA COLLECTION

The collection of spoken language corpora is an important research area and represents a significant portion 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[4]. Similarly, progress in spoken language understanding is closely linked to the availability of spoken language corpora.

Since September 1994, we have greatly expanded our data collection effort and record subjects on a regular basis. The recordings are made in office environement, simultaneously with a close-talking, noise cancelling Shure SM10 and a table-top Crown PCC160 microphone. We now collect data directly using up-to-date versions of our spoken language systems for l'Atis and Mask. The cumulative number of subjects and queries recorded are shown in Table I. As of Feb95, 163 speakers had recorded 8611 queries. We are collecting speech at a rate of over 1000 queries per month from at least 20 speakers. There are an average 10 words per query. The total number of words, and the number of distinct words are also shown. The new word rate decreased from Sep. though Dec. In mid December a new version of the l'Atis data collection system was installed. This version corrected some problems in the maintenance of the dialog history which had caused the subjects to repeat several times the same query, and integrated a new version of the speech recognizer. The combined improvements changed significantly the user's interaction with the system. With the more performant speech recognizer, speakers speak more easily and use longer and more varied sentences. This also leads to the occurrence of more new words in the queries. They are also more

Corpus	*#Sents*	*WAcc*	*NL*	*SLS*
L'ATIS *Oct94*	225	90.2%	85%	84%
L'ATIS *Jan95*	225	93.7%	89%	88%
MASK *Jan95*	205	78.0%	85%	60%
MASK *Mar95*	205	83.2%	91%	75%

TABLE II. Results for L'ATIS and MASK.

likely to perceive that the recognition errors are their fault, rather than the system's. As a result they continue to speak relatively naturally to the system, enabling us to record more representative spontaneous speech which will in turn be used to improve the system. We started recording MASK data in January and thus far have recorded 77 speakers for a total of 3876 queries. The average sentence length is 8 words, shorter than for L'ATIS. This difference is probably linked to the performance of the speech recognizer which for now has a higher word error than for L'ATIS due to the limited amount of training data. There are 753 distinct words in the MASK training queries, with only 271 which are not in the L'ATIS training data. As more data is recorded we will be able to improve the performance of the MASK data collection system as was observed for L'ATIS.

7. EXPERIMENTAL RESULTS

The spoken language system has been evaluated on 225 queries from 10 speakers for L'ATIS and 205 queries from 10 speakers for MASK. On L'ATIS, the speech recognition word accuracy on L'ATIS has been improved from 90.2% in Oct94 to 93.7% in Jan95. This improvement is directly linked to the availability of additional training data, particularly for the LM. The NL component was also evaluated on the exact transcriptions of the same set of spoken queries, without removing spontaneous speech effects such as hesitations or repetitions. A correct caseframe instanciation is obtained for 89% of the queries.

The speech recognition word accuracy is much lower for MASK but has improved from 78% in Jan95 to 83% in Mar95. NL understanding has increased from 84% to 91%. This improvement is mainly due to handling reductions, a concept unknown to the initial system.

For the Jan95 L'ATIS system we observe that there is only a slight degradation in performance for the complete spoken language system relative to the NL component. Since the performance of the speech recognizer for MASK has not yet reached the same level as that for L'ATIS, a larger degradation occurs.

Two frequent errors in understanding are common to both tasks. The first involves 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. The second is 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.

One of the recent improvements made to the maintenance of the dialog history for L'ATIS is the capability of relaxing previously specified constraints. This allows us to treat requests such as *"une autre compagnie" (another company)* or *"tous les vols" (all the flights)*. In the first case, all constraints on the company and the flight-number are removed. This capability accounts for half of the improvement between the Oct94 and Jan95 systems.

8. SUMMARY

We have presented our recent activities in developing spoken language systems for 3 applications in the travel domain: L'ATIS, MASK, and RAILTEL, and prototype systems currently in use 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. Working simultaneously on several tasks with the same core system allows a broader view of the problems, and the opportunity to quickly integrate improvements made in one task to the other tasks .

REFERENCES

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

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

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

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

[5] S.M. Katz, "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer," *IEEE Trans. ASSP*, **35**(3), 1987.

[6] L.F. Lamel, J.L. Gauvain, B. Prouts, C. Bouhier, R. Boesch, "Generation and Synthesis of Broadcast Messages," *ESCA Workshop on Applications of Speech Technology*, Sep. 1993.

[7] P. Price, "Evaluation of Spoken Language Systems: The ATIS Domain," *Proc. DARPA Workshop on Speech and Natural Language*, Hidden Valley, PA, 1990.