The Spoken Language Component of the Mask Kiosk

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

Abstract

The aim of the Multimodal-Multimedia Automated Service Kiosk (MASK) project is to pave the way for more advanced public service applications by user interfaces employing multimodal, multi-media input and output. The project has analyzed the technological requirements in the context of users and the tasks they perform in carrying out travel enquiries, and developed a prototype information kiosk that will be installed in the Gare St. Lazare in Paris. The kiosk will improve the effectiveness of such services by enabling interaction through the coordinated use of multimodal inputs (speech and touch) and multimedia output (sound, video, text, and graphics) and in doing so create the opportunity for new public services. Vocal input is managed by a spoken language system, which aims to provide a natural interface between the user and the computer through the use of simple and natural dialogs. In this paper the architecture and the capabilities of the spoken language system are described, with emphasis on the speaker-independent, large vocabulary continuous speech recognizer, the natural language component (including semantic analysis and dialog management), and the response generator. We also describe our data collection and evaluation activities which are crucial to system development.

1 Introduction

Information technology has the potential to improve information and services for the general public. However, often such services fail to realize their potential and are frequently underused. The problems with public service provision are exemplified in the kiosks currently available for rail travellers to obtain information about train services and local facilities, and to purchase tickets. The average transaction time at such kiosks is four times as long as with service staff. As a consequence, the kiosks are under-utilised, being used primarily at night when no agent is present, or when there are long lines for human service. Evidently the technology being used does not meet the particular needs of intended users in the context of the tasks they want to perform. Specifically, the rigidity of touch-screen-based, menu-driven user-interfaces prevents users transacting tasks fluently, as when communicating with another person.

In the ESPRIT Multimodal-Multimedia Automated Service Kiosk (MASK) project the goal is to develop a more advanced interface employing multimodal, multi-media input and output so as to pave the way for more advanced public services. In the context of the

project, the technological requirements needs of users carrying out travel enquiry tasks have been analyzed and a prototype information kiosk has been developed. The kiosk, which will be installed in the Gare St. Lazare in Paris for evaluation with real users, should improve the effectiveness of such services by enabling interaction through the coordinated use of multimodal inputs (speech and touch) and multimedia output (sound, video, text, and graphics) and in doing so create the opportunity for new public services. The partners in the MASK project are MORS (coordinator, F), SNCF (F), LIMSI-CNRS (F), and UCL (UK).

The role of LIMSI in the project is to develop the spoken language component of the MASK kiosk. Spoken language systems aim to provide a natural interface between a user and a computer by using simple and natural dialogs to enable the user to access stored information. The main information provided by the MASK kiosk 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. In the next section an overview of the architecture of the MASK spoken language component is given, followed by sections detailing the subcomponents of the overall system. This is followed Section 3 concerned with our data collection activities, which represent a significant portion of the effort in system development. In Section 4 we describe the objective and subjective evalutation measures used to assess progress.

2 System Overview

An overview of the spoken language system for information retrieval is shown in Figure 1. 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

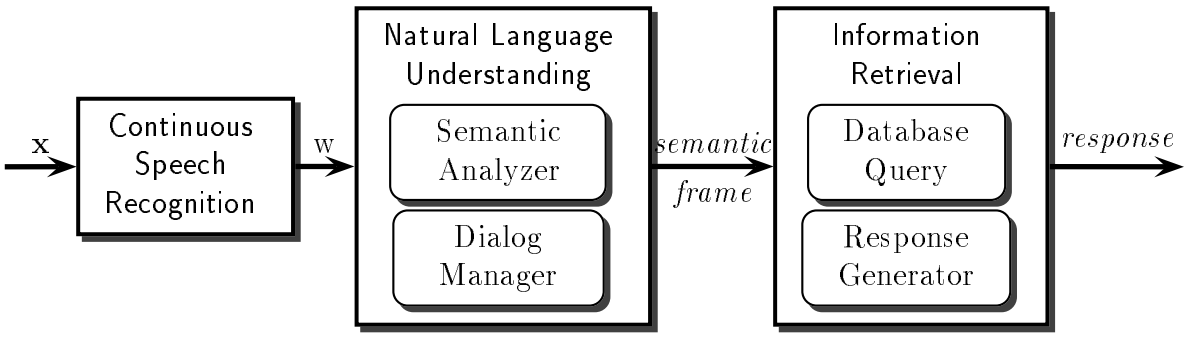

Figure 1: Overview of the spoken language information retrieval system. **x** is the input speech signal, **w** is the word sequence output by the speech recognizer.

on the languages known to the system in order to achieve the best possible performance.

2.1 Speech Recognizer

Speech recognition is concerned with the problem of transcribing the speech signal as a sequence of words. Today's most performant systems are for the most part based on a statistical modelisation of the talker. From this point of view, message generation is represented by a language model which provides estimates of $\Pr(w)$ for all word strings w, and the acoustic channel encoding the message w in the signal x is represented by a probability density function $f(x|w)$. The speech decoding problem consists then of maximizing the a posteriori probability of w, or equivalently, maximizing the product $\Pr(w)f(x|w)$.

This formulation highlights the main problems to resolve: that of estimating the language model $\Pr(w)$, and the acoustic encoding $f(x|w)$. Language modeling entails incorporating constraints on the allowable sequences of words which form a sentence. Statistical n-gram models attempt to capture the syntactic and semantic constraints by estimating the frequencies of sequences of n words. A backoff mechanism[9] is used to smooth the estimates of the probabilities of rare n-grams by relying on a lower order n-gram when there is insufficient training data, and to provide a means of modeling unobserved n-grams. The n-gram statistics are estimated on the orthographic transcriptions of the training set of spoken queries. Word classes are used for lexical items such as the cities, days, months, are used to provide more robust estimates of the n-gram probabilities, when there is no reason to believe that differences in their frequencies in the training data are significant or representative. Sub-language models are being developed for use with system directed dialogs. These sub-languages will allow the search space to be reduced, improving the accuracy and reducing the computational needs. The recognition lexicon has on the order of 1500 entries, including 500 station/city names, and is represented phonemically with a set of 35 phonemes.

Acoustic modeling makes use of continuous density hidden Markov model (HMM) with Gaussian mixture. Context-dependent phone models are used to account for allophonic variation observed in different contextual environments.

The speech recognizer is a software-only system (written in ANSI C) that runs in what is close enough to be perceived as real-time on a standard RISC processor. The system is speaker-independent, so that no speaker-specific enrollment data is needed for a new user. Speaker independence is achieved by using acoustic models which have been trained on speech data from a large number of representative speakers, covering a wide variety of accents and voice qualities.

Two types of acoustic compensation are used in the recognizer to account for the background acoustic noise and acoustic channel variability. The first compensates for the typical noise present in the acoustic environment of the train station by explicitly modeling different typical noise levels. Between user transactions, the system periodically assesses the level of the background noise and chooses the most appropriate acoustic model set. The second type accounts for differences in the particular acoustic channel for the current utterance by the user. These differences which arise from the differences in the speakers' heights, distance

Semantic category	*Example*
train-time	*Quels sont les **horaires** des trains allant de Paris à Lyon ?*
	What are the **times** of trains from Paris to Lyon ?
fare	*Quel est le **prix** du billet ?*
	How **much** is the ticket ?
change	*Quels sont les **changements** ?*
	What are the **correspondences** ?
type	*Quel est le **type** du train qui arrive à 20 heures 5 ?*
	What **type** of train is the one arriving at 20:05 ?
reserve	*Je veux **réserver** une place dans le train de 8 heures 10.*
	I want to **reserve** a seat on the 8:10 train.
service	*Quelles sont les **prestations** offertes dans ces trains ?*
	What **services** are available on these trains ?
reduction	*Qu'est-ce qu'un billet **Jocker** ?*
	What is a reduction **Jocker** ?

Figure 2: MASK concepts.

and orientation relative to the microphone, as well as speaker-specific vocal characteristics are minimized by use of incremental cepstral mean removal.

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.

2.2 Natural Language Understanding

The natural language component is concerned with understanding the meaning of the spoken query. This component has two subcomponents - semantic analysis and dialog management. The semantic analyzer carries out a caseframe analysis[5, 4] to determine the meaning of the query[3, 1], 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 and the appropriate keywords. This undertaking, which is quite important (and difficult), is obviously task-dependent but hopefully language independent. However, in transfering to another task in a related domain (such as for air travel information and reservation) many of the same concepts and keywords are conserved[11]. The concepts for the MASK task as shown in Figure 2 are train-time, fare, change, type, reserve, service and reduction and have been determined by analysis of queries taken from the training corpora to augment the *a priori* task knowledge.

The caseframe parser has been implemented in C++. The caseframe grammar is described in a declarative file so as to allow for easy modification of the cases. Casemarkers are surface indicators designating a case and provide syntactic constraints necessary to ex-

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)
```

Figure 3: Example semantic frame.

tract the meaning of the request. For example, in "**de** Paris **à** Marseille", the preposition **de** designates Paris to be the departure city and the preposition **à** designates Marseille to be the arrival city. In the phrase "à 14 **heures**", **heures** is an example of a postmarker, designating 14 to be a time. Since the understanding of numbers is very relevant to the travel information task (appearing in times, 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. The query "*Montrez-moi les trains de demain matin allant à Marseille en provenance de Paris avec un changement à Lyon.* (*Show me trains for tomorrow morning going to Marseille from Paris with a change in Lyon.*)" will result in the same caseframe.

The dialog manager ensures the smooth interface between the user and the computer. The dialog process formally consists of transitions between five dialog states: opening formalities, information, stagnation, confirmation subdialogs and closing formalities[2]. 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.

2.3 Information Retrieval

The response generator uses the semantic frame to generate a database request to the database management system. The retrieved information is reformatted for presentation to the user along with an accompanying natural language response. A vocal response is optionally provided along with the written and tabular information. The generation of responses is complex bacause if too much information is given, it may be difficult for the user to extract the important part. If not enough information is returned, the interaction will take longer, as the user will need to ask for more detailed or additional information. In the MASK project we are experimenting with different forms of response - text strings, tables, and ticket images, so as to facillitate the transfer of information to the user.

When vocal feedback is provided the speech must 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. Therefore simple playback of pre-recorded speech is used for fixed messages that are unlikely to be changed. However, since is not possible to present variable information using direct playback of pre-recoreded speech, we make use of a speech concatenation approach[10] 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.

3 Data Collection

In order to develop a spoken language system, task-specific data must be available to train the acoustic and language models. Collection of such spoken language data represents a significant portion of the work in system development. The use of additional acoustic and language model training data has been shown to almost systematically improve performance in continuous speech recognition[6]. Similarly, progress in understanding is closely linked to the availability of spoken language corpora.

Using the MASK spoken language system, we have recorded over 10,000 queries from over 150 speakers. The recordings are made at LIMSI in office environement, simultaneously with a close-talking, noise cancelling Shure SM10 and a table-top Crown PCC160 microphone. Each subject participates in a 2 hour recording session, during which time they solve at least 10 MASK scenarios with a range of complexities. Two example scenarios are given in Figure 4. The scenarios are periodically modified to ellicit a wider variety of vocabulary items, such as city names, dates and times of travel. We also include specific scenarios in which users need to find out information about concepts not yet handled by the system, to see how they react in order to help us develop ways to detect such situations and to guide the user accordingly. To help assess our progress in system development, at the end of the recording session each subject completes a questionnaire addressing the user-friendliness, reliability, ease-of-use of the MASK data collection system.

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. Politeness forms also occur in about 3% of the sentences: *please* (1.5%), *hello* (1.5%), *thank you* (0.5%). Other interjections such as *then*, *well*, and *okay* occur in about 3% of the utterances. Spontaneous speech phenomena such as hesitations, false starts and reparations occur in about 25% of the queries. The filler word *euh* occurs in 9.4% of the queries, and breath noises (inspiration and expiration) were marked in about 11% of the transcriptions.

(S1) *You want to go from Grenoble to Paris next Friday as late as possible. You want to take a direct TGV and to pay a reduced fare. Reserve a non-smoking place.*
(S2) *You are traveling from Bordeaux to Avignon next Sunday. You have a reduction* ***Carissimo.*** *Your dog is traveling with you. Reserve an aisle seat in a second class, smoking car. Will you need to change trains?*

Figure 4: Example scenarios used for data collection.

Month	*Jan*	*Feb*	*Mar*	*Apr*	*May*	*June*	*July*
#speakers	12	42	78	106	113	143	153
#queries	208	1603	3825	6219	6853	9587	10368
total #words	1.6k	12.3k	29.1k	44.5k	48.6k	69.6k	77.8k
#distinct words	273	737	975	1120	1168	1349	1444
#new words	-	420	211	113	34		68

Table 1: MASK data collection status.

The MASK spoken language system uses a mixed-initiative dialog strategy, where the user is free to ask any question, at any time. However, in order to aid the user to obtain a reservation, the system prompts the user for any missing information needed for database access. There are on average 14 queries to solve a scenario. Approximately one-third of the system responses are direct requests asking the user to provide specific information. These direct requests involve the class of travel (27%), date (23%), departure city (16%), time (15%), smoking (13%), and arrival city (6%).

650 dialogs were analyzed to see how subjects respond to system initiatives. Subjects provided a direct response to these requests over 60% of the time, however they frequently also provided additional information. For example, when asked for the departure city, the user often also specified the arrival city and/or the time of travel. In collaboration with the SNCF, have recently carried out recordings at the Gare St. Lazare in Paris in order to have access to a more realistic potential user population. During a 10 day period, over 120 subjects solved an average of 2.5 scenarios using the current prototype system. We are in the process of transcribing and classifying the queries, and analyzing the dialogs.

4 Evaluation

The development of a spoken language system is incremental, where errors are analysed and the system is refined. The link between development and evaluation is so tight that we consider evaluation to be part of the development process. Periodic evaluation on specified

Corpus	*#Sents*	*WAcc*	*NL*	*SLS*
MASK *Jan95*	205	78%	85%	60%
MASK *Apr95*	205	85%	93%	79%
MASK *Aug95*	205	90%	93%	85%

Table 2: Evaluation of the MASK spoken language system.

test sets allows us to continually monitor progress through objective performance measures.

4.1 Objective Evaluation Measures

The speech recognizer is evaluated in terms of speed and recognition accuracy (word and sentence error). An analysis of the recognition errors is carried out to determine their effect on the understanding performance. The understanding component is evaluated using typed versions of the exact transcriptions of spoken queries including all spontaneous speech effects, such as hesitations or repetitions, (so as to evaluate the understanding component without intrusion of errors made by the speech recognizer) and on the recognized word string. A semi-automatic method is used which compares the resulting semantic frame to reference semantic frames for each test query.

In Table 2 gives evaluation results of the MASK data collection system on a set of 205 queries from 10 speakers. The word accuracy is has improved from 78% in Jan95 to 85% in Apr95, to 90 in Aug95. 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 85%.

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 *"Montrez-moi tous les trains" (show me all the trains)* after having specified a departure time.

4.2 User Evaluation

It is also important to assess the overall performance of the system from the point of view of the subjects. Since March'95 all subjects have completed a questionnaire (Figure 5) addressing the user-friendliness (1-3), reliability (4-6), ease-of-use (7-9) of the MASK system. Subjects are also asked what are the good aspects of the system, how it should be improved,

and if they would use such a potential system. Information about the subject includes how often they travel by train, how they obtain their tickets, and their computer experience.

1. Is it easy to speak to the system?
2. Is the system easy to understand?
3. Does the system respond fast enough?
4. Are you confident in the information given by the system?
5. Did you get the information you wanted?
6. Are you satisfied with the information?
7. Did the system recognize what you said?
8. Did the system understand what you said?
9. If the system did not understand you, was it easy to reformulate your question?

Figure 5: User questionnaire.

(a)	*Experience*	*Ease-of-use*	*Reliability*	*Friendliness*
	Questions	*1 2 3*	*4 5 6*	*7 8 9*
	Expert (17)	7.7	7.5	7.1
	Novice (44)	7.2	6.5	5.7

(b)	*Age*	*Ease-of-use*	*Reliability*	*Friendliness*
	<24 (31)	8.3	6.9	6.6
	24-50 (25)	7.3	6.7	6.0
	>50 (5)	6.3	6.9	5.8

(c)	*Travel*	*Ease-of-use*	*Reliability*	*Friendliness*
	Frequent (19)	6.9	5.9	6.1
	Infrequent (25)	7.5	7.0	6.1

Figure 6: User reponses on the ease-of-use, reliability, and friendliness of the MASK spoken language system.

The results of the analysis of the responses of 61 speakers are shown in Figure 6 an a scale of 10. In (a) the responses are divided based on the comfort of users with the system, independent of their age and their travel habits. Users were classed as novices if they had difficulties speaking with the system or using the computer. In general "expert" users (no difficulty speaking with the system and used to working with computers) were more at ease with the system, and judged it to be more user-friendly, easier to use, and more reliable than the novices. The novices were more likely to critique the reliability of information obtained

from the system, whereas the experts criticized problems in understanding or dialog. In Figure 6(b) the responses are subdivided by the age of the subjects, where there is a clear tendancy of younger subjects to assess the system more favorably than the older subjects. As shown in (c) for the naive subjects, frequent train travelers are slightly more sceptical and dissatisfied with the system than infrequent travelers. In general, users express an interest in using such types of systems, and often ask to come back to participate in future experiments.

5 Summary

This paper has described the spoken language component of the MASK kiosk. A prototype spoken language system has been used to collect data at LIMSI and at the SNCF, and this data has been used for system development. On a test set of 205 queries from 10 speakers, the speech recognition word accuracy is 90%. Natural language understanding of typed transcriptions of the same queries is 93%. The complete spoken language system has an understanding rate of 85%. We expect that as more data is collected the understanding rate will improve, as we previously observed for our L'ATIS system[11]. 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 speak more easily and use longer and more varied sentences. They are also more likely to perceive that errors are their own 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. We have recently collected data from 120 subjects at the Gare St. Lazare in Paris. In early 1996, the MASK kiosk will be installed in the Gare St. Lazare. Data collected on-site will be used to further improve the system, better matching the system's capabilities to the user's needs.

References

[1] S.K. Bennacef, H. Bonneau-Maynard, J.L. Gauvain, L. Lamel, W. Minker, " A Spoken Language System For Information Retrieval," *Proc. ICSLP'94*, Yokohama, Japan, Sept. 1994.

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

[3] H. Bonneau-Maynard, J.L. Gauvain, D. Goodine, L. Lamel, J. Polifroni, S. Seneff, "A French Version of the MIT-ATIS System: Portability Issues," *Proc. Eurospeech'93*, Berlin, Germany, Sept. 1994.

[4] B. Bruce, "Case Systems for Natural Language," *Artificial Intelligence*, **6**, 1975.

[5] Ch.J. Fillmore, "The case for case," in *Universals in Linguistic Theory*, Emmon Bach & Robert T. Harms (eds.), Holt, Rinehart and Winston, Inc., 1968.

[6] 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," *Proc. IEEE ICASSP-94*, Adelaide, Australia, April 1994.

[7] J.L. Gauvain, L.F. Lamel, G. Adda, M. Adda-Decker, "Continuous Speech Dictation in French," *ICSLP-94*, Yokohama, Japan, Sept. 1994.

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

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

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

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

[12] L. Lamel, S. Rosset, S. Bennacef, H. Bonneau-Maynard, L. Devillers, J.L. Gauvain, "Development of Spoken Language Corpora for Travel Information", *Eurospeech'95*, Madrid, Spain, Sept. 1995.

[13] H.Ney, "The Use of a One-Stage Dynamic Programming Algorithm for Connected Word Recognition," *IEEE Trans. ASSP*, **32**(2), 1984.