Spoken Language Dialog System Development and Evaluation at LIMSI

Lori Lamel
Spoken Language Processing Group
LIMSI-CNRS, BP 133
91403 Orsay cedex, France
lamel@limsi.fr http://www.limsi.fr/tlp

ABSTRACT

The development of natural spoken language dialog systems requires expertise in multiple domains, including speech recognition, natural spoken language understanding and generation, dialog managment and speech synthesis. In this paper I report on our experience at LIMSI in the design, development and evaluation of spoken language dialog systems for information retrieval tasks. Drawing upon our experience in this area, I attempt to highlight some aspects of the design process, such as the use of general and task-specific knowledge sources, the need for an iterative development cycle, and some of the difficulties related to evaluation of development progress.

1. INTRODUCTION

At LIMSI we have experience in developing several spoken language dialog systems for information retrieval tasks[5, 11, 16, 19, 1]. Our recent activities in this area have been mainly in the context of European projects, such as ESPRIT MASK, Language Engineering RAILTEL and ARISE, Tide HOME-AOM, Esprit LTR Concerted Action DISC, and a French language action launched by the AUPELF-UREF.

In this paper I provide an overview of our spoken language dialog system, describing the main components as well as some modifications for specific tasks. Most of the examples will be drawn from our train travel information systems for the MASK and ARISE projects. The spoken language system integrates a speaker-independent continuous speech recognizer (based on HMM with statistical language models), a semantic analyzer (based on a caseframe grammar) and a dialog manager. The dialog manager is the central controller of the entire system as it manages contextual understanding, the dialog history, information retrieval and response generation. The dialog management aspect of the system has become more important as we have gained experience with spoken language dialog systems.

In our view, spoken language systems should provide a natural, user-friendly interface with the computer, allowing easy access to the stored information. Our goal is to obtain high dialog success rates with a very open dialog structure, where the user is free to ask any question or to provide any information at any point in time. Our basic dialog strategy (described in [3]) has been significantly modified as a result of user trials (described in [16] and [19]) in order to adhere to some generic dialog guidelines. To improve performance within this open dialog strategy, we make use of implicit confirmation (using the caller's wording to the extent possible) and change to a more constrained dialog level when the dialog is not going well.

Our first experience with a spoken language dialog system (SLDS) was developing a French version[5, 2] of ATIS (Air Travel Information Service) a designated common task for data collection and evaluation within the ARPA Speech and Natural Language Program[23]. This work was initialized in collaboration with the MIT-LCS Spoken Language Systems Group, and the natural language understanding (NLU) component of the MIT ATIS system[24] was ported to French[5]. The SLDS was ported to a train travel information retrieval task in the context of the ESPRIT Multimodal Multimedia Service Kiosk (MASK) project, aiming to develop an innovative, user-friendly prototype information kiosk combining tactile and vocal input[11, 7, 18]. The MASK interface has a self-presentation illustrating the use of the kiosk and explaining the different types of transactions available; an intuitive interface with easy switching between tasks (such as information or ticketing); a facial image of a clerk to let the user know what the system is doing; and a two-level help facility with fixed time-outs.

The same basic SLDS technology was adapted to a prototype telephone service in the context of the LE-MLAP RAILTEL (Railway Telephone Information Service)[3, 16] and LE-3 ARISE (Automatic Railway Information Systems for Europe) projects[19]. In the ARISE system for main intercity connections Callers are able to obtain information taken from the French Railways (SNCF) static timetables and additional information about services offered on the trains, fares and fare-related restrictions and reductions. A prototype French/English service for the high speed trains between Paris and London is also under development. In the context of the AUPELF-UREF ac-

tion B2, different dialog strategies are being explored with the PARIS-SITI spoken dialog system providing tourist information[4, 8].

2. SYSTEM OVERVIEW

An overview of the spoken language system architecture is shown in Figure 1. The main components for spoken language understanding are the speech recognizer, the semantic analyzer, and the dialog manager, which controls the information retrieval component including database access and response generation. For the MASK system which also allows tactile input, there are the multimedia interface and the touch screen. The speech recognizer is a medium vocabulary ($\sim$ 2000 words), real-time, speaker-independent, continuous speech recognizer which transforms the acoustic signal into the most probable word sequence. The recognizer output is passed to the semantic analyzer which extracts the meaning of the spoken query using a caseframe analysis [2]. Semantic interpretation is carried out in two steps, first a literal understanding of the query, and then its reinterpretation in the context of the ongoing dialog. The mixed-initiative dialog manager, which has the goal of providing information to the user, ensures communication between the user and the DBMS. The dialog manager maintains both the dialog and generation histories. The generation component outputs a natural language response based on the dialog state, the caller's query, and the information returned from database access. Information can be returned to the user in the form of synthesized speech or visually if a display is available. Natural-sounding utterances are synthesized by concatenation of variable-sized speech units stored in the form of a dictionary[15].

The ability to interrupt the system (a barge-in capability) is often considered to be important for usability. Adding this capacity required modifications to several modules. Firstly, recording and speech recognition must be active at all times, even when the system is synthesizing a response. Software-based echo cancellation, applied to the recorded signal using the known synthesized signal, is used to suppress the system response. If speech is detected, or if there is a tactile input, synthesis is stopped. There are dialog situations in which barge-in is disabled to ensure that the caller hears the entire message.

3. DATA COLLECTION

For SLSs it is necessary to collect application-specific data, which is useful for accurate modeling at different levels (acoustic, lexical, syntactic and semantic). Data collection 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[13]. Similarly, progress in spoken language dialog systems is closely linked to the availability of spoken language corpora. Acquiring sufficient amounts of text training data is more challenging than obtaining acoustic data. With 10k queries relatively robust acoustic models can be trained, but these queries contain only on the order of 100k words, which probably yield an incomplete coverage of the task (ie. they are not sufficient for word list development) and are insufficient for training *n*-gram language models.

MASK	*#Subjects*	*#Queries*	*#Words*	*#Distinct*
Jun95	146	9.6k	69.6k	1180
Dec95	313	18.7k	150.8k	1690
May96	392	26.6k	205.4k	2060
Jun97	478	50.6k	351.2k	2560

ARISE	*#Calls*	*#Queries*	*#Words*	*#Distinct*
Aug97	2787	36.4k	179.7k	2529
Dec97	6130	84.5k	412.3k	3677
Mar98	6545	88.4k	436.3k	3764
Oct98	10262	149.1k	663.3k	4610

Table 1: Data collection for the MASK and ARISE systems. Word fragments are not counted.

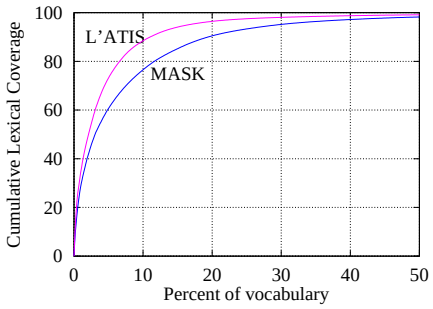

Figure 2: Percentage of transcription covered as a function of the percentage of words.

It is common practice to use a WOz setup or a bootstrap system to collect an initial corpus. The bootstrap system is often based on prior work: acoustic models or training data may be taken from a different task; an initial vocabulary can be obtained by considering the task and introspection; and a simple language model can be estimated on a set of typed queries. These queries can also be used to develop an initial set of rules for the semantic analyzer. Our experience is that as the system improves, subjects speak more easily and use longer and more varied sentences. This leads to the occurrence of more new words and new formulations in the queries. Table 1 summarizes the cummulative data collected for MASK and ARISE with different system versions. The number of distinct words found in the corpora is relatively small compared to the total number of words. The lexical coverage as a function of word frequency is shown in Figure 2 for MASK and L'ATIS data.

4. SPEECH RECOGNIZER

The speech recognizer is a software-only system that runs in real-time on a standard RISC processor. Some of the design issues in developing a speech recognizer for an

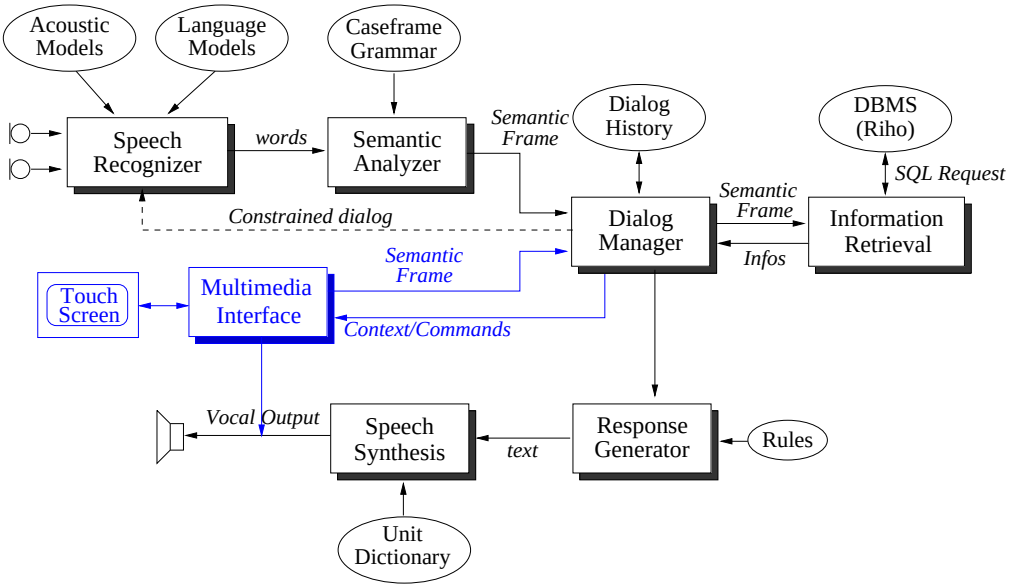

Figure 1: Spoken language system architecture.

SLDS are discussed in [10] and [12]. Statistical models are used at the acoustic and word levels. Acoustic modeling makes use of continuous density hidden Markov model (HMM) with Gaussian mixture. 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. Context-dependent phone models can be used to account for allophonic variation observed in different contextual environments. The word recognition graph is built by putting together word models according to the grammar in one large HMM. A pronunciation graph is associated with each word so as to allow for alternate pronunciations, including optional phones. Our recognition lexicon is represented phonemically. Specific phone symbols and acoustic models can be used to model non-speech events such as silence, breath noise and hesitations.

N-gram backoff language models[14] are estimated on the orthographic transcriptions of the training set of spoken queries, with word classes for cities, dates and numbers providing more robust estimates of the *n*-gram probabilities. It is fairly common practice to use compound words for common word sequences that are subject to strong reduction and coarticulation in spontaneous speech.

For dictation tasks, it is relatively easy to select a recognition vocabulary as text data are generally available. In contrast, for SLDSs generally only very limited (if any) transcribed data available for lexical and language modeling. The word list is usually designed using *a priori* task-specific knowledge and completed by task-specific collected and transcribed data. For example, the recognition vocabulary of the MASK and ARISE systems contains about 2000 words, including 600 station names selected to cover the French Railway commercial needs, other task-specific words (dates, times), and all words occurring at least twice in the training data. For spontaneous speech, it is important that the lexicon include pseudo words for hesitations "ehr", and extraneous filler words. Breath noise is also often modeled as a separate lexical item.

In order to reduce the number of understanding errors due to speech recognition, a confidence score is associated with each hypothesized word. If this score is below an empirically determined threshold, the hypothesized word is marked as uncertain. These uncertain words can be ignored by the understanding component or used by the dialog manager to start clarification subdialogs. On average, rejection tends to lead to a longer dialog, since some correct words are ignored. However with the use of rejection, the overall dialog success rate is improved. Station names can be optionally spelled so as to support improved recognition performance with a large number of cities, as this is critical for the task. In our current implementation the speech recognizer outputs the best word sequence with a confidence score. It is also able to provide a word lattice.

5. SEMANTIC ANALYSIS

The text string output by the recognizer is passed to the semantic analyzer. This component first carries out literal understanding of the recognizer output, and then reinterprets the query in the context of the ongoing dialog. In literal understanding, the semantic analyzer applies a case-frame grammar to determine the meaning of the query, and builds an appropriate semantic frame representation[2]. 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 this component is in defining the concepts that are meaningful for the task and the appropriate keywords. The

concepts needed to carry out the main ticketing task (for both MASK and ARISE) concern train times, connections, fares and reservations (including reductions and other constraints). Other concepts are used to handle general information available about reductions and services. The concepts have been determined using *a priori* information about the task and have been completed by an analysis of the queries in the training corpora.

Contextual understanding consists of reinterpreting the utterance in the context of the ongoing dialog, taking into account common sense and task domain knowledge. The semantic frame resulting from the literal understanding is reinterpreted using default value rules, and qualitative values are transformed into quantitative ones. For example, if the departure month has not been specified *"I would like to leave on the 4th"*, the current month is taken by default (or the next month if the 4th has already past). The semantic frame corresponding to the current utterance is then completed using the dialog state and history in order to take into account all the information previously given by the user, as well as the questions posed by the system.

Although the understanding component of our current SLDSs make use of the caseframe grammar, at LIMSI we have been exploring the use of statistical approaches for this component[22]. The attraction of statistical methods stems from their success in speech recognition, and their ability to model unseen formulations, with human intervention being limited to labeling (or correcting labels). Known disadvantages are that stochastic models require large training corpora in order to reliably estimate model parameters, and that being estimated on training data, common events are better modeled than rare ones. Also, generalizations that can be made relatively easily by humans may not be automatically learned.

6. DIALOG MANAGEMENT

Dialog management is very challenging in the context of natural, mixed-initiative systems where the user is free to change the direction of the dialog at essentially any point in time. In order to be closer to a real dialog situation, representatives from LIMSI and VECSYS visited the Douai SNCF Information Service to observe how the human-human dialogs are performed and what strategies are used by the human operators.

The main objectives of the dialog strategy are:

1) To never let the user get lost. The user must always be informed of what the system has understood. This is of particular importance as most users are unfamiliar with talking to a machine.

2) To answer directly to user questions. The system responses should be as accurate as possible and provide immediate feedback of what was understood.

3) To give to the user the opportunity, at each step, to correct the system. This capability is needed to be able to correct for recognition errors, but also to let the user correct what s/he said or to have a change of mind.

4) To avoid misunderstanding. Even though users are able to correct the system at any moment, we have observed that they tend to not do so. It is therefore important to minimize recognition errors, as users can not be expected to correct the system. This is our motivation for rejecting unreliable hypotheses.

For MASK the interaction of the multimedia interface and the spoken language system is via the dialog manager. The multimedia interface interprets tactile commands and generates a semantic frame compatible with the SLDS. The dialog manager integrates the tactile information into the current dialog context and controls database access. The high-level decisions are taken by the dialog manager based on the context and the state of the interface, and low-level presentation decisions are taken directly by the multimedia interface. An important difference in dialog strategies is offered by the input modes. The tactile strategy is a command driven dialog, where the user must input specific information in order to move on to the next step. Vocal input allows a real mixed-initiative dialog between the user and the system, where the user can guide the interaction or be guided by the system via the help messages.

After contextual understanding, the dialog manager either uses the semantic frame to generate an SQL-like request to the database management system or prompts the user to fill in missing information. If the result of contextual understanding is void, or if it is in contradiction with the dialog context, the system can ask the user to repeat (either directly "I'm sorry, I did not get that, can you please repeat it?" or indirectly "Excuse me?"). For MASK and ARISE the user is required to specify four key items before accessing the database (a static copy of the SNCF database: the departure and arrival stations, the date and approximate time of travel. The day and time can be specified exactly (March 14th) or in a relative manner, such as *next Monday, early morning, late tomorrow afternoon*. Interpretative and history management rules are applied prior to generation of the DBMS request. These rules are used to determine if the query contains new information, and if so, if this information is contradictory with what the system has previously understood. If a contradiction is detected, the dialog manager may choose to keep the original information, replace it with the new information, or enter into a confirmation or clarification subdialog. Post-processing rules, which take into account the dialog history and the content of the most recent query, are used to interpret the returned information prior to presentation to the user.

Constraint relaxation is used in retrieving timetable information in order to provide a more cooperative dialogue and response. For example, if no train satisfies the user's request, the system relaxes constraints on the departure time in order to find the closest train before or after the specified time. In this case it is important that the system response is justified by informing the user that the proposed train is the

closest match to their request. If not, the user may assume that the system has made a mistake.

The dialog strategy has undergone (and is still undergoing) significant changes as we gather more experience with a wider range of users. For example, in our initial RAILTEL system[3] we decided to return information for up to 3 trains. If more trains satisfied the user's request, the system returned the number of trains in the time period and the departure times of the first and last train. In the current ARISE system, only one train is proposed, that which is closest to the request. If the user specifies a time range (e.g., early morning), the train closest to the middle of the specified time is returned. The user is able to ask for a different train (the preceding/following one, an earlier/later train, the first/last train, a direct train, etc). For the telephone system we have found this approach effective in reducing the overall dialog duration (This is similar to what a human agent does). In contrast, for the information kiosk, it is easy to display a list of trains and let the user choose one of them.

However, it is evident that no one single dialog strategy will satisfy all users, as different users need differ amounts of guidance, and there will be differences in performance of the speech recognizer and semantic analyzer across users. In order to improve performance, a two-level dialog strategy has been implemented for ARISE where a system-directed dialog is entered if a problem is detected in obtaining departure and arrival station names or the date of travel[19]. When the constrained dialog is active, the speech recognizer makes use of a dialog state dependent language model. A constrained dialog can be initiated by the system if the user does not respond to the system prompt for one of the four basic items (timeout), or in cases where the information received by the system is contradictory with what was previously understood. Such constrained dialogs apply only to the departure and arrival cities, and the travel date. For example, if the system understood a change in the departure or arrival city, one of the following strategies is used depending upon the state of the dialog: the system may choose to ignore the information; it can ask for an explicit confirmation of the new city; or it can ask the user to repeat the information. If the caller changes one of these items during the confirmation request, implicit confirmation is used in the following prompt. The directiveness of the prompt increases if the user does not supply the requested information, suggesting for example that the caller spell the city name.

The generation component converts a generation semantic frame into a natural language response. The form of the natural language response depends on the dialog context, and whether or not the same information was already presented to the user. We aim to give a direct response to the caller, highlighting the new information and directly integrating the information given in the user's request. This immediate feedback allows the user to know what the system has understood[19]. If the user does not change the information items, they are considered as implicitly confirmed. Careful attention has been paid to construction of sentences that contain the appropriate information and to the generation of natural-sounding utterances[3]. We try to use short responses, so as to keep the caller in tighter contact with the system, and to make for a more natural dialog.

Aspect	*Measures*
Speech recognizer	word error, content word error, confidence measures
Semantic analyzer	semantic frame error, slot error
Dialog	response
System	global measures (success, #turns, time, waiting time...)
Subjective	questionnaires

Table 2: Some assessment metrics for spoken language dialog systems.

7. EVALUATION

While there are commonly used measures and methodologies for evaluating speech recognizers, the evaluation of spoken language systems is considerably more complicated due to the interactive nature and the human perception of the performance. It is therefore important to assess not only the individual system components, but the overall system performance using objective and subjective measures. Evaluation plays an integral role in system development, which we consider as an ongoing activity. Different types of evaluation can be used, each with their particular strengths and costs. In general, it is advantagous when the evaluation can be carried out automatically, which requires labeling of the test data. This type of evaluation can be applied to individual system components, particularly the speech recognizer and the semantic understanding component. A multilevel error analysis can be used to distinguish between errors due to a particular component and those propagating from preceding stages[16]. Table 2 summarizes the evaluation aspects and measures discussed in this paper. When experimenting with new user interfaces[20] and dialog strategies, it is often useful to carry out an informal assessment of system performance and capabilities and how these are perceived by users.

An important concern is obtaining realistic user trials. These are obviously needed to properly evaluate the prototype or potential service, but can be risky if done too prematurely. Being a research laboratory we are not developing commercial systems and as a consequence usually do not often have access to the final user. However, we would like our user trials to be as realistic as possible. As a consequence, we recruit subjects on ongoing basis to provide data for system development and evaluation. For the MASK project over 600 users were recruited to test different versions of the system, both at our laboratory at a Parisian train station. In addition to this data, periodic evaluations were

carried out by UCL and SNCF with different system versions, prior to the final evaluation with 200 subjects[18]. For ARISE we have recorded over 10000 calls, with a total of 149k queries. Three rounds of evaluation were carried by the SNCF to assess usability and performance of different versions of the system.

For the SNCF tests, subjects were recruited by a hostess at a Parisian train station. The subjects were asked to test a new, experimental automatic ticket kiosk (MASK) or telephone service (ARISE), and were given a gift certificate for their participation. Subjects carried out 3 or 4 scenarios, and completed a short questionnaire after each call and estimated the completion time. After the final scenarios subjects completed a more in depth questionnaire, which asked general questions about the subject and their computer experience and travel habits, in addition to specific questions about different aspects of the prototype system.

7.1. Component Assessment

While from the viewpoint of the user, only the global performance measures are important, it is important for the system developers to look closely at the different sources of errors within each component of the complete system. In our evaluation work we have focused on the speech recognizer, semantic analyzer, and dialog manager components, and have not paid much attention to the information retrieval or synthesis components.

Evaluation of the speech recognizers and speech synthesizers have been the subject of numerous, long term activities. While there are well known tests to assess speech synthesizers, these have not been widely used in the context of SLDSs. Most systems make use of available text-to-speech systems or use synthesis by concatenation. The former has the advantage of being able to pronounce any text, at the cost of naturalness. Synthesis by concatenation requires that all speech units are prerecorded, and changing the prompt usually requires carrying out new recordings.

For speech recognition, the most commonly used metric is the word error rate:

$$100 * \frac{\textit{\#Substitutions} + \textit{\#Insertions} + \textit{\#Deletions}}{\textit{Total Number of Reference Words}}$$

For read speech, the reference text is known, i.e. words are "defined" (given an agreed upon tokenization). For spontaneous speech it can be difficult to agree on the reference string. Contractions are common (*what's / what is, he's / he is / he has, dunno / I don't know / I do not know*, as are hesitations/fillers *(uhm, hmmm, uh-huh)*, non-speech events *(breath, sniffles, cough, throat clearing)*, and word fragments and mispronunciations *(fr-, *district)*. While these can also be found in read speech, they are much less common than in spontaneous speech. Related activities which have been investigated are the relevance of measures, scoring of word fragments, and phonological scoring[9], and the use of confidence measures[6]. In some cases relatively low word errors have been reported for speech recognizers of information retrieval systems, particularly for the ARPA ATIS task. Some indicative word error rates are shown in Table 3. However, these numbers can be misleading as the word error measures all differences between the exact orthographic of the query and the recognizer output. Many recognition errors (such as masculine/feminine forms, or plurals) are not important for understanding. As can be observed for the MASK word error rates, there is often a substantial degradation in performance when moving from laboratory recruited subjects to more representative user populations.

Task	*Vocabulary Size (words)*	*Word Error*
ATIS	∼1500 (11/46 cities)	2-14% laboratory data
MASK	∼2000 (600 stations)	7% laboratory data 13% kiosk
ARISE	∼2000 (600-1000 stations)	10-20% telephone

Table 3: Some indicative word error rates for SLDSs.

Error	*dep-city*	*arr-city*	*dep-time*	*arr-time*	*dep-date*
#slots	78	80	216	95	86
Reco	5.2%	4.2%	18.5%	8.2%	29.6%
Und	3.6%	4.4%	7.0%	0.5%	6.0%

Table 4: Recognition and understanding error rates on semantic slots for the MASK system.

Methodologies have been proposed to evaluate the semantic analysis[23, 21]. This evaluation can be carried out on the speech recognizer output, or on typed versions of the exact transcriptions of spoken queries including all spontaneous speech effects, such as hesitations or repetitions, (so as to evaluate this component without intrusion of errors made by the speech recognizer). In order to evaluate the semantic analysis component, we make use of semi-automatic method which compares the resulting semantic frame to a reference semantic frame. For each slot which is incorrectly instanciated, the error source, recognition or understanding, is marked. It is then straightforward to compute the incorrect slot instanciation rate (recognition or understanding) for the semantic frame by simply dividing the number slot errors by the total number of slots. We can consider the slot error rate to correspond to the word error rate on content words, and the semantic frame error to the sentence error. Recognition and understanding slot error rates on a set of 368 MASK transactions are shown in Table 4 for the *departure-city*, *arrival-city*, *departure-time*, *arrival-time* and *departure-date*. The error rates correspond to the number of erroneous slots divided by the total number of slots for each type. The average query recognition error on this data was 16.2% and the understanding error 5.4%, illustrating that recognition errors do not necessarily entail understanding errors. Similarly, not all understanding errors are important for dialog success, for example, interpreting the time period as "around 10 pm"

instead of "after 10 pm" may not affect the information obtained from the database, and therefore has no effect on the dialog. It has been our observation that such minor understanding errors pass unobserved by the user, whereas more important understanding errors will lead to longer dialogs, as the user tries to correct the error.

We assess the dialog to determine if it was successful by looking at the system responses. Knowing both the correct transcription of the spoken query, the recognizer hypothesis and the semantic frame, we can determine the error source. The dialog error is calculated as the ratio of incorrect responses and the total number of system responses. The dialog error in obtaining timetable information was 16% on 58 calls recorded with the ARISE system during a two-day test period last June[19]. Reservations, which require specifying the class of travel, seating preference and reduction, had a failure rate of 11%. A higher error rate (30%) was obtained for diverse questions, due in part to functionality limitations. Since knowing when a dialog has finished is a difficult task, we analyzed how the dialogs ended. 12% of the dialogs ended without a closing formality (ie. the caller hung up) without saying goodbye. Such abrupt endings can occur when a caller got the desired information, or because the user was frustrated. We also analyzed the use of barge-in on the same data. Users interrupted the system in 72% (42) of the calls, speaking during 13% (122 of 958) system responses. When barge-in was observed during a call, it was used on average to interrupt 3 system responses. Barge-in was observed in a variety of contexts, but was most (40% of the interruptions) often used to respond to questions before they were finished. For example, when the system is uncertain about a station name, the caller is prompted to say and optionally spell the city name. (*Give your departure city and spell it if you like. For example, Paris, P A R I S.*) 25% of the barge-ins seemed to be inadvertant. The caller was seeminly engrossed in their thoughts, talking to the system and unaware that the system was responding. In contrast to our expectations, barge-in was only rarely used (6% of the cases) to correct the system, and usually to change the date of travel.

7.2. Global Performance Assessment

Global evalation measures concern the entire user interaction, and include both objective and subjective measures, as well as external observations. Some objective measures are the transaction completion and success rates, the total duration of the interaction, the waiting time, the number of dialog turns, the number of repetitions, corrections and interruptions. In the case of failure, ie. the user obtained the wrong information, or did not receive any information, the stage of failure may be noted. Subjective user assessemnt usually addressess qualitative criteria such as the ease of use, perceived speed, and perceived reliability. The effectiveness of speech can be compared and combined with other modalities, such as touch screen or keypad for input and a visual display for output.

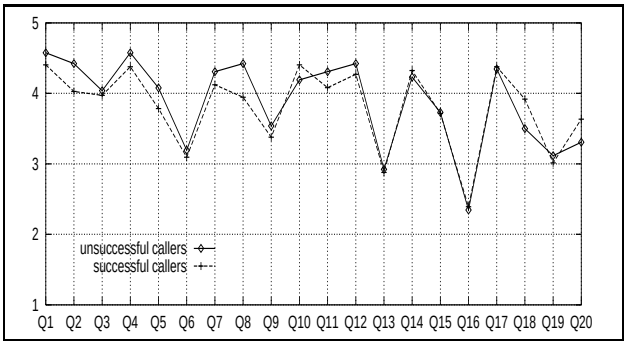

Figure 3: RAILTEL usability profile: results for successful and unsuccessful callers. Q1: ease-of-use, Q2: confusing, Q3: friendliness, Q4: complex, Q5: use again, Q6: reliability, Q7: control, Q8: concentration, Q9: efficiency, Q10: fluster, Q11: too fast, Q12: stress, Q13: prefer human service, Q14: complicated, Q15: enjoyable, Q16: needs improvement, Q17: politeness, Q18: information obtained, Q19: faster than human, Q20: understood

Ease-of-use
1. Is it easy to speak to the system?
2. Is the system easy to understand?
3. Does the system respond fast enough?

Reliability
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?

Friendliness
7. Did the system recognize what you said?
8. Did the system understand what you said?
9. If you were not understood, was it easy to reformulate your question?

Figure 4: MASK questionnaire.

In the RAILTEL project a common questionnaire was designed and used to assess the usability of the three prototypes. The questionnaire contained 20 statements, with which users were asked if they agreed (on a scale of 1 to 5). For example, the statement Q1 is "I found the system easy to use." and the statement Q13 is "The system was faster than a human." Figure 3 shows the user assessements of the LIMSI system as a function of the success of the call[16]. We can see that there is very little difference in ratings for successful and unsuccessful callers. This may be an undesired side-effect of the evaluation process, in that the subjects who participated are carrying out a scenario, and do not really care about the returned information. They therefore may assess the system in a different manner than a real user. This point highlights the importance of continual, ongoing evaluation with more and more realistic users.

User questionnaires can be relatively short, addressing the user's perception of the transaction and system, or quite detailed. A questionnaire we used with MASK subjects (Figure 4) addressed three main issues: ease-of-use, reliability and friendliness[11]. For ARISE we now use a simple ques-

The ARISE service is easy to use.
I got the information I wanted.
The system seemed to understand me.
I understood the system.
I found the responses too long.

Figure 5: ARISE questionnaire.

tionnaire, shown in Figure 5, which is completed after each call. In addition, the subjects are asked to write down what they asked for and what information the system returned to them. In this way we are able to verify whether or not the system really gave the desired information.

8. DISCUSSION & PERSPECTIVES

Enabling efficient, yet user-friendly interaction for access to stored information by is quite difficult. Most existing services are directive, restricting what the caller can ask at any given point in the dialog, and limiting the form of the request. Some laboratory prototypes allow a more open, user-initiated dialog, but performance is generally lower than what can be obtained with more restricted dialog stuctures. Developing and evaluating spoken language dialog systems is complicated due to the interactive nature and the human perception of the performance. It is also time-consuming as much of the analysis must be carried out manually. It is important to assess not only the individual system components, but the overall system performance using objective and subjective measures.

SLDSs must recognize spontaneous speech, which is usually produced by the talker who is speaking while composing the message. Spontaneous speech is known to have variations in speaking rate, speech disfluencies (hesitations, restarts, incomplete words or fragments, repeated words) and rearranging of word sequences or "incorrect" syntactic structures[25]. Subsequent system modules must be able to deal with both the structures of spontaneous speech and recognition errors. By associating confidence scores with each hypothesized word the semantic analyzer and dialog modules can choose to ignore uncertain items, that could be misrecognitions. Although such rejection may lead to a longer dialog, since some correct words are ignored, the overall dialog success rate can be improved.

From the dialog perspective, it is important that the user is aware of what the system has or has not understood. Close communication is thus essential for dialog success. This coupling can be obtained by using immediate feedback of what was understood, combined with implicit confirmation. Explicit confirmation may be used when the system is uncertain or has understood contradictory information.

Considering the overall system development, human factors should be taken into account early in the design phase, and in successive modifications. The MASK propotype kiosk was developed after analysis of the technological requirements in the context of users and the tasks they perform in carrying out travel enquiries[20]. Our preliminary analysis of a barge-in capability for the ARISE system (which *a priori* was considered to be very important for usability, at least for directive systems) indicates that it is not heavily used, and is not used in the manner we had anticipated (i.e., to correct misrecognized items). This may be partially due to the experimental conditions, as callers do not really need the information they are asking for, and therefore may not notice (or care about) the errors.

An important issue that was highlighted during the SNCF user trials is that users do not distinguish the functionalities of the service from the system responses. Even if a system is able to detect some out-of-functionality requests, and responded that it is unable to handle these, such responses are not satisfactory for users. For example, if the user wants to reserve for several people and the system informs him/her that it is unable to reserve for more than one person at a time, this is logical and correct from the spoken language system developer's point of view, who considers the dialog to be a success . The user, however, has not obtained what s/he desired, and may not be satisfied with the response.

User trials of the MASK kiosk carried out with over 200 subjects demonstrated that for this task multimodality is more efficient (faster and easier) than unimodality as some actions are better carried out by voice and others by touch. These studies also showed that subjects performed their tasks more efficiently as they became familiarized with the MASK system, learning to exploit the vocal input and benefiting from the multiple modalities. 74% of the users claimed to have never or rarely encountered difficulties in using the system, and 98% were largely satisfied with the usability and simplicity of use.

9. ACKNOWLEDGEMENTS

This paper is based on collaborative work with colleagues at LIMSI (in particular Samir Bennacef, Jean-Luc Gauvain and Sophie Rosset) as well as at the SNCF and the Vecsys Company. We thank the SNCF for providing the information database RIHO for use in the ARISE project and for their contribution to assessment by carrying out the user trials. We also thank the Vecsys company for their contributions to the generation and synthesis modules. Our work in spoken language system development and evaluation has benefited from partial support from the following projects: ESPRIT MASK, Language Engineering MLAP RAILTEL and LE-3 ARISE, ESPRIT-LTR Concerted Action DISC, TIDE HOME-AOM, AUPELF-UREF ARC B2.

10. REFERENCES

1. J. Shao, N.-E. Tazine, L. Lamel, B. Prouts, S. Schroeter, "An Open System Architecture for Multimedia and Multimodal User Interface," *3rd TIDE Congress*, Helsinki, June, 1998.

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

3. S. Bennacef, L. Devillers, S. Rosset, L. Lamel, "Dialog in the RAILTEL Telephone-Based System," *ICSLP'96*, Philadelphia, pp. 550-553, Oct. 1996.
4. H. Bonneau-Maynard, L. Devillers, " Dialog Strategies in a tourist information spoken dialog system," *Specom'98*, St. Petersburg, Oct. 1998.
5. 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*, Berlin, Sept. 1993.
6. L. Chase, "Word and Acoustic Confidence Annotation for Large Vocabulary Speech Recognizers," to appear in *Computer, Speech & Language*.
7. H. Dartigues, F. Bernard, A. Guidon, J.N. Temem, "The MASK project : new passenger service kiosk technology," *World Congress on Railway Research '97*, Florence, pp. 513-518, Nov. 1997.
8. L. Devillers, H.Bonneau-Maynard , "Evaluation of Dialog Strategies for a Tourist Information Retrieval System," *ICSLP'98*, Sydney, Dec. 1998.
9. W. Fisher, J. Fiscus, A. Martin, D. Pallett, M. Przybocki, "Further Studies in Phonological Scoring," *ARPA Spoken Language Systems Technology Workshop*, Austin, TX, pp. 181-186, Jan. 1995.
10. J.L. Gauvain, J.J. Gangolf, L. Lamel, "Speech Recognition for an Information Kiosk," *ICSLP'96*, Philadelphia, pp. 849–852, Oct. 1996.
11. J.L. Gauvain et al. S. Bennacef, L. Devillers, L. Lamel, R. Rosset, "Spoken Language component of the MASK Kiosk" in K. Varghese, S. Pfleger(Eds.) "Human Comfort and security of information systems", Springer-Verlag, 1997.
12. J.L. Gauvain, L. Lamel, "Large Vocabulary Continuous Speech Recognition: from Laboratory Systems towards Real-World Applications," *Institute of Electronics, Information and Communication Engineers*, J79-D-II:2005–2021, Dec. 1996.
13. J.L. Gauvain, L. Lamel, G. Adda, M. Adda-Decker, "The LIMSI Continuous Speech Dictation System: Evaluation on the ARPA WSJ Task," *ICASSP-94*, Adelaide, **1**, pp. 557-560, April 1994.
14. S.M. Katz, "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer," *IEEE Trans. ASSP*, **35**(3), 1987.
15. L. Lamel, J.L. Gauvain, B. Prouts, C. Bouhier, R. Boesch, "Generation and Synthesis of Broadcast Messages," *ESCA-NATO Workshop on Applications of Speech Technology*, Lautrach, Germany, pp. 207-210, Sept. 1993.
16. L. Lamel, S.K. Bennacef, S. Rosset, L. Devillers, S. Foukia, J.J. Gangolf, J.L. Gauvain, "The LIMSI RailTel System: Field trials of a Telephone Service for Rail Travel Information," *Speech Communication* **23**, pp. 67-82, Oct. 1997.
17. L. Lamel, S. Rosset, S. Bennacef, H. Bonneau-Maynard, L. Devillers, J.L. Gauvain, "Development of Spoken Language Corpora for Travel Information," *Eurospeech'95*, **3**, pp. 1961-1964, Madrid, Sept. 1995.
18. L. Lamel, S. Bennacef, J.L. Gauvain, H. Dartigues, J.N. Temem, "User Evaluation of the MASK Kiosk," *ICSLP'98*, Sydney, Dec. 1998.
19. L. Lamel, S. Rosset, J.L. Gauvain, S. Bennacef, "The LIMSI ARISE System," *IVTTA'98*, Torino, pp. 209-214, Sept. 1998.
20. A. Life, I. Salter, J.N. Temem, F. Bernard, S. Rosset, S. Bennacef, L. Lamel, "Data Collection for the MASK Kiosk: WOz vs Prototype System," *ICSLP'96*, Philadelphia, pp. 1672–1675, Oct. 1996.
21. W. Minker, "Evaluation Methodologies for Interactive Speech Systems," *LREC'98*, Granada, pp. 199–206, May 1998.
22. W. Minker, "Stochastic versus rule-based understanding for information retrieval," *Speech Communication*, **25**(4), pp. 223-247, Sep. 1998.
23. P. Price, "Evaluation of Spoken Language Systems: The ATIS Domain," *DARPA Workshop on Speech & Natural Language*, Hidden Valley, PA, 1990.
24. S. Seneff, "TINA: A Natural Language System for Spoken Language Applications," *Computational Linguistics*, **18**(1), pp. 61-86, 1992.
25. A. Stolcke, E. Shriberg, "Statistical Language Modeling for Speech Disfluencies, *ICASSP-96*, Atlanta, GA, **I**, pp. 405-408, May 1996.