

A Probabilistic Approach to Person-Robot Dialogue

Renato De Mori^{1,2}, Jacqueline Bourdeau², and Roland Kuhn¹

¹ McGill University School of Computer Science, Montreal, Canada H3A 2K6

² Centre de Recherche en Informatique de Montreal, Montreal, Canada H3T 1P1

Abstract

This paper surveys current work on systems that support spoken man-machine dialogue. Subsequently, we describe our current research on a natural language interface for robot control, indicating how our system uses probability scores to evaluate and combine hypotheses generated at different conceptual levels.

Keywords

oral dialogue - speech recognition - person-machine dialogue
- natural language interface - probabilistic approach

1 Introduction

Oral dialogue (OD) means a spoken conversation with initiative shared between the user and computer. OD focuses on spoken communication on both sides of the conversation, either in a system whose sole mode of interaction is speech, or combined with visual information, editing, pointing or other direct manipulation. The concept is being explored, and experimental work is underway with dialogue systems which make use of oral dialogue.

Dialogue systems do not imply a specific type of input/output interface, whether spoken or written. They may accept a combination of the two, where each plays a role. Natural Language (NL) techniques are used to complement, confirm, validate or correct the spoken input. Since the dialogue may include information from another source of

input, such as visual information, dialogue systems often imply multimodal interaction, and have to be designed and managed with this in mind.

Oral dialogue systems do not imply any particular type of processing. They may employ HMMs, parsing techniques, other techniques, or combinations of these. The current trend is to find a complementary role for different techniques, involving much methodological work on such questions.

Oral Dialogue has a strong, two-way relationship with Speech Recognition (SR). SR research considers a large variety of topics, including studies in signal processing and connectionist methods. The goals are to improve the recognition performance rate, and to solve problems posed by multiple speakers and large vocabularies. Areas addressed in human-machine communication include speed, performance, and accuracy, as well as exploration of the application of SR to real world situations. SR techniques form the basis for OD; however, knowledge gained through dialogue system research can lead to improvement of SR performance.

Oral dialogue design can be conducted using different approaches: modelling human-human dialogue, simulating a machine, or by directly building a working prototype in order to test and subsequently refine a design. One major design issue is to find optimal trade-offs between user-friendliness and robustness, between user freedom and performance.

Oral dialogues are different from other kinds of NL phenomena in that they have a structure, are goal-oriented, and the choice of words is constrained. Dialogue understanding is more tractable than spontaneous speech since the concepts and the words are well defined, the domain is restricted, and the goal is well known. Each phrase has a specific function in the conversation, such as information gathering, answer, interpretation, etc.

2 Overview of Person-Machine Dialogue Systems

Since the 70's, research work has progressively evolved from speech recognition to speech understanding, leading researchers to consider the NL understanding approach and techniques. This is how SR researchers started to explore problems common to speech and language processing, and to consider parsing techniques and the linguistic approach with its classical dimensions: syntactic, semantic, lexical and pragmatic. Some of them [3,12] refer to Searle's Speech Act Theory. Most people working on the integration of SR and NL language understanding techniques have a background in SR, and are interested in dialogue constraints. The use of dialogue constraints varies from correcting SR errors to selecting word candidates and reducing the search space.

Artificial Intelligence (AI) also plays a role in this evolution. Some researchers simply apply established techniques; others use a more global AI conceptual approach. Most recently, researchers have begun to use knowledge acquisition techniques for dialogue systems rather than relying only on traditional knowledge-based techniques.

As much as the desire to apply SR technology, it seems that it is the research on applications that led the researchers to work directly on dialogue systems. These applications include machine translation [16], telephone services [12], database inquiry [2,38], office work automation [1,25], telerobotics [4], air traffic control [18]. Design issues are raised for driving voice-driven dialogues [11,25]. Conceptual issues are suggested for multimodal dialogue [33,1], and intelligent dialogue [7,8], aiming at, ideally, their integration.

Towards the end of the 80's, the domains of natural language processing and speech understanding began to intersect, and dialogue systems became a new research field.

Below is a description of the activities of the research teams whose research focuses on oral dialogue systems, as well as their particular approaches and some problems that arise out of this new research field.

2.1 Speech Group, Medialab, MIT, USA

Medialab Speech Group's projects are to some extent influenced by the industrial orientation of this lab, as well as by the futuristic vision of its head, N. Negroponte. In his view, speech could become the dominant mode of interaction because it is frequently the most available communication channel, and because it allows the user to do remote work. The two goals of the group's approach are: 1) to integrate speech and language understanding, 2) to consider speech as one of several concurrent channels, across which a message content can be redundant. Bolt's research objective [1] is to obtain an integrated multimodal interface, where the design issue is to orchestrate the different modes, in order to obtain improved usability, performance and robustness. Principles for such orchestration may include: reducing the user's mental load, and obtaining summation and redundancy of information through the use of several modes.

2.2 Laboratory for Computer Science, MIT, USA

Spoken language understanding is part of the program of the Spoken Language Systems Group, as an extension of the SUMMIT SR system developed there. TINA is a natural language system developed by Seneff [28] specifically for speech understanding, in order to be integrated into SUMMIT. The system works with context-free grammars, ATNs, and lexical functional grammars. In converting context-free grammar rules into a probabilistic network structure, the grammar includes arc probabilities reflecting the frequency of occurrence of the syntactic structures. These probabilities are trained automatically from a set of parsed sentences; training can happen only if parsing was successful.

Zue and Seneff [42] developed VOYAGER as an application in order to test the two components, SUMMIT and TINA, connected in a serial manner. SUMMIT converts the speech signal into word hypotheses, and TINA gives an interpretation, which is then transformed into query functions. A back-end is responsible for response generation. Semantics was incorporated into TINA's parse tree. VOYAGER gives directional assistance for the Cambridge area. An experiment was conducted to test components and overall system response. The system could not parse 66.1% of the sentences, due to recognition errors, unknown words, or unseen linguistic structures. Experts found that 97% of the functions generated were correct. Recognition rate can be improved by incorporating context-dependent phoneme models.

2.3 Carnegie-Mellon University, USA

Under the direction of R. Reddy, work is being done by K. F. Lee and others on SR techniques; Ward focuses on speech understanding while Rudnický concentrates on dialogue system design and experimentation. S. R. Young, in collaboration with Ward and Hauptmann, developed a system using dialogue context in SR [39,40]. The MINDS

system generates expectations in order to predict the upcoming dialogue, then organizes constraints into hierarchical sets called “layers” to reduce its search space. Each prediction is then translated into a grammar usable by the speech recognizer. MINDS derives predictions from pragmatic knowledge sources such as dialogue structure, task semantics, general world knowledge, and user knowledge. Instead of using dialogue constraints to correct recognition errors or to select word candidates, MINDS uses layered contextual constraints to circumscribe dynamically the search space for words in a speech signal. It guides the recognition process from the very beginning. Experiments with CMU’s recognizer, SPHINX, showed effective reduction of search space, as well as improved recognition performance. Performance is higher with users whose behaviour is consistent.

Ward [36] examines problems specific to spontaneous speech, and solutions that make SR systems robust enough to be able to handle phenomena such as filled pauses, restarts, interjections, ellipsis and ungrammatical constructions, problems common to SR and NL. A combination of techniques from both fields is developed, namely modelling acoustic properties (instead of words), with a flexible control structure. Grammar is used as a local constraint to group words into phrases, while global constraints from semantics combine the sequence of meanings into a defined action. A system called PHOENIX has been developed using the HMM word models of SPHINX.

Rudnický [25] raises different design issues, beginning their investigation with task language analysis. To him, constraining the language means modelling the task language by user observation, in order to get a language as natural as possible; he calls it a “habitable” language. Fluent interaction is the second design objective, which implies modelling the dynamics of dialogue, as well as user characteristics. Error repair facilities and allowing the user to define new words are other design features which make the whole environment habitable. Experiments conducted on a financial planning task, with a voice-driven interface to the SPHINX system [25], comparing voice only, voice and keyboard, and keyboard only interfaces, demonstrated good performance in terms of task completion. However, subjects using voice took more time, this being due to two SR limitations: lack of speed and of accuracy.

2.4 Cambridge University

Work done by S.J. Young with Proctor [38] at Cambridge University, within an Alvey project called VODIS (Voice Operated Database Inquiry System), deals with design and implementation issues. The main focus is on intelligent dialogue control, in order to compensate better for SR errors. In the view of the authors, using a finite state transition network would rapidly become combinatorially too large, i.e. intractable. Their design methodology is to build a dialogue controller using an object-oriented framework, in which objects contain information needed to complete a transaction. With the inheritance mechanism of frames, dialogue design is data-driven, and is realized through successive refinement. As experiment shows, this methodology is valid and also easier to test, since each topic can be tested individually.

2.5 CNET-IRISA, France

Similarly to MIT and CMU, researchers at CNET-IRISA started from their SR system, KEAL, to explore understanding and dialogue systems. Siroux and Gillet [29] first intro-

duced syntactic, semantic and pragmatic knowledge to implement dialogue function as an interface to KEAL. A context-free grammar takes into account semantic features, which are then interpreted to represent information relevant to the subsequent dialog. The dialogue manager instantiates a model graph, makes predictions, and controls the question-answer type of oral exchange. Experimental work was done in the PALABRE ESPRIT project by Guyomard and Siroux [12] on a telephone service application, the French general information service called "Yellow Pages". Dialogue modelling was performed after observation, in order to structure and constrain a cooperative type of dialogue. Dialogue features include cooperation, follow-on dialogues, use of implicit utterances, and instructions given to users. Experimentation showed the validity of this realistic approach in terms of cooperation and prediction; however, predictive power is a detriment to user's freedom. Linguistic expertise facilitates the task of recognition by reducing the number of utterances and their length. New requests issued from follow-on dialogue can be classified into three categories: slidings, mutations, and reproductions.

2.6 CRIN-INRIA, France

At CRIN-INRIA, Pierrel and Haton work on speech understanding; Haton's work on labeling is part of the SAM ESPRIT project. Carbonell and Pierrel [4] have been working since the beginning of the 80's on dialogue systems design. Their approach is knowledge-based, and they propose an architecture based on four knowledge sources, accessed independently by the dialogue component, giving as much modularity as possible. These knowledge sources are: acoustic-phonetic, prosodic, syntactic-semantic and lexical. The dialogue controller builds up a dialogue history out of dynamic information on the dialogue process. Knowledge about dialogue structures and strategies is represented by plans, which are then selected by the dialogue manager according to dialog progress. Souvay et al. [30] have applied an implementation of this design to the oral command of a Sonar system. Voice input was introduced to replace a complex input system composed of keyboard and menus, trackball and switches. Keyboard and mouse could be used in case of deficiencies in the voice input. Results are positive in terms of usability and performance: operators can keep looking at the screen and concentrate a command in a few words. This, with attendant reduction in mental overload and overall workload, allows operators to work faster.

2.7 LIMSI, France

At LIMSI, Matrouf et al. [18,19] have developed a task-oriented and knowledge-based approach to build a dialogue system for air traffic controller training. After having analyzed and categorized the corpus of messages used in real life, they have deduced rules and criteria, and built a model based on the theory of frames. The dialogue system consists of four parts: message analysis, validity control, correction and interpretation. Knowledge sources are vocabulary, phraseology, task model, and dialogue history. The focus is on the representation of the language model in order to reduce the search space. Pragmatic knowledge is used to make predictions which dynamically limit the recognition search space. The dialogue module is made of a network of frames. After message analysis, other modules are invoked for error detection and correction, message generation, updating of the dialogue network, and prediction for the next message. Predictions are

made from current context and dialogue history, using a probabilistic syntax, in which the probability of the occurrence of a concept is given a weight. Prediction is then propagated to the arc probabilities of the automaton corresponding to the concept. For experimentation, the AMADEUS SR system made at LIMSI was adapted in order to use transition probabilities in the recognition process. Results indicated that the use of probabilities and semantic knowledge increased the message recognition accuracy to 96%, and that dynamic probabilities provided an additional 0.5% improvement. Combined use of dynamic probabilities and semantic knowledge-based dialogue strategies gave a performance of 96.5% at the sentence level.

2.8 Advanced Telecommunications Research Institute, Japan

Integration of language and speech processing is extensively researched at ATR, as related to machine translation of telephone dialogues [16]. The goal is an automatic telephone interpretation system that will transform a spoken dialogue from the speaker's language to that of the listener automatically and simultaneously. Such spoken dialogues include elliptic and anaphoric expressions, as well as syntactically ill-formed expressions, all of which must be taken into account. A more limited goal is to handle task-oriented areas, where domain knowledge is expected to disambiguate the spoken language. Linguistically-based constraints of syntactic, semantic, and pragmatic information are utilized to narrow the number of word candidates. A method is developed that reduces the number of candidates presented by the SR component. In this method, the relationship of dependency is used for the analysis of spoken Japanese utterances. The current approach to analyzing Japanese dialogues is based on a lexico-syntactic grammar framework in terms of feature structures.

2.9 Bolt, Beranek and Newman, USA

At BBN, Schwartz and Makhoul have developed a SR system called BYBLOS using HMMs with context-dependent phoneme modelling, and are working on parallel implementation of the system. Roucos [24] developed a spoken language system, adding NL sources to BYBLOS: a unification grammar for syntax, a higher-order intensional logic language for semantics, and a Montague Grammar style framework for interfacing syntax and semantics. HMMs are used to determine a word lattice; then, a lattice parser finds all possible word sequences. The semantic interpreter determines the highest scoring meaningful ones, and interprets them. Experimental results with the DARPA Task Resource Management database demonstrate an improvement of accuracy with syntax from 71% to 87%, and, with both syntax and semantics, to 92%.

2.10 SRI, USA

SRI's Speech Research Group, under Murveit and Bernstein, aims at creating real-time systems for interactive problem-solving, by integrating SR and NL understanding. The key idea is to develop a parser that incrementally generates a state transition network to replace a finite state grammar in their HMM-based SR system [21]. Finite state constraints are used in order to permit integration into SRI's SR system, DECIPHER. The approach is to generate a dynamic grammar network that represents NL knowledge

in a state transition network; only portions of the network within a pruned search are expanded. This approach reduces computation time while maintaining a very tractable recognition search space.

2.11 University of Erlangen, Federal Republic of Germany

At the University of Erlangen, Niemann and collaborators have developed one of the earliest speech understanding and dialogue systems, EVAR [22], where linguistic knowledge is used to interpret a sentence and discover the user's intent. Semantic and pragmatic knowledge are combined to filter hypotheses at different steps of the understanding process. Ehrlich and Niemann [9] use semantic and pragmatic knowledge to reduce the number of hypotheses by analyzing complex syntactic constituents. Their approach to semantic analysis is based on the caseframe theory and deep case theory. Results from experiments show that the number of hypotheses could be reduced by half through this filtering. Sagerer's approach to SR systems [27] is knowledge-based, using semantic networks.

This concludes our survey of leading research activity in Speech Research Groups. Despite the diversity of the work described, one may discern certain common themes. All of the groups aim to enhance recognition by incorporating a degree of understanding. In order to achieve this, they look for ways to make full use of their existing SR systems, adding linguistic processing techniques, while not reducing performance or robustness. The most common approach is to introduce NL constraints and/or AI techniques; much methodological work is being done on experimenting with different techniques. The NL techniques vary depending on the SR methodology, on the type of architecture of the existing SR system, and also on the overall orientation of the lab. The role of the understanding component, as compared with that for recognition, varies from reducing errors, or correcting them, to reducing the search space. Experiments performed with these spoken language systems, either on standard speech databases, or on new applications, show an improved recognition capacity. In some cases, limitations are reported to be due to the basic SR system.

In addition to the research programs in speech labs, one should be aware of other work related to OD, either on linguistic and communication problems in natural dialogue by Bunt [2,3], user modelling by Wahlster [35], or dialogue ergonomics by Waterworth [37]. We briefly introduce conceptual work conducted at DCIEM (Defense and Civil Institute of Environmental Medicine) by Taylor, and by Edwards and Mason. Taylor [31,32] proposes the Layered Protocol Model for analyzing and designing human-computer interaction. This view is inspired by psychological principles and also by previous work on ISO-OSI protocols. Taylor suggests the metaphor of computer-computer communication to describe human-computer dialog, and the Layered Protocol Model as a structure to design dialogue systems. In this view, multimodal dialogue consists of multiple dialogs conducted over multiplexed channels, each having its own coding and feedback features. Redundancy is necessary, as is feedback without delay, in order to obtain effective communication without misunderstanding. Edwards and Mason [7,8,17] focus on the aspect of intelligence in dialogue and propose a goal-oriented dialogue design model.

Psychological views like these, as well as linguistic ones, must be taken into account in OD systems design and evaluation, to make sure that psychological principles are respected, and in order to be able to discriminate between computing problems and other kinds of problems e.g. linguistic, psychological, and so on.

3 Person-Robot Dialogue

This section reports on progress in the construction of a system that supports spoken man-machine dialogue. A natural language interface for robot control has been designed. Such a system must meet the following, extremely demanding, criteria:

1. It must have access to a model of the environment, so that sentences or phrases that refer to the environment can be disambiguated
2. It must have a model of the robot's capabilities, so that at any point in the dialogue it knows what the robot is currently doing, what it is currently capable of doing, and what it has done in the recent past (note that this kind of state knowledge is only useful in the presence of the environment model just mentioned)
3. It must have a dialogue model, enabling it to manage a dialogue with the user - that is, it must be capable of pursuing its own conversational goals (the primary goal will usually be that of determining exactly what the user wants the robot to do) while responding flexibly to conversational initiatives by the user
4. At lower levels of the linguistic hierarchy, it must recognize sequences of words uttered by the user and convert them into semantic tokens which are passed to higher levels
5. It must permit information transfer between system components, so that information gained by one of the components facilitates the task of others.

The last point is particularly important. In our system, information passed from one component to another is often expressed in terms of **probabilities**. For instance, consider the way in which the speech recognition component interacts with the other components. The most urgent problem in large-vocabulary speech recognition is constraining the size of the search space. Thus, high level knowledge about the environment, semantics, discourse structure, user goals, and dialogue history is applied during spoken man-machine interaction to provide constraints for the speech recognition component. Since this component uses probabilistic models to evaluate competing acoustic hypotheses, the higher-level components provide it with probabilistic predictions about what will be said next. This uniform representation enables the speech recognition component to combine information from different levels in a consistent way. It may be the case that there is more than one plausible semantic interpretation of what the user said - if so, each of the interpretations will be passed up to higher levels of the system, along with the set of corresponding probabilities. Thus both the input and the output of this component are expressed probabilistically.

From a software engineering point of view, this approach leads to a transparent system, with a modular internal structure. There is no need of complicated protocols to decide how to assess information provided by other components - competing hypotheses can be scored in terms of their probabilities, allowing a best-first search to take place. If the history of speech recognition is any indication, a probabilistic system is also likely to be extremely robust.

4 Robot Behaviour and the Structure of Person-Robot Dialogue

Consider the circumstances in which person-robot dialogue will take place. As stated above, the system managing the natural-language robot-user dialogue requires three models: an **environment model**, a **robot state model**, and a **dialogue model**.

The **environment model** contains the system's knowledge about the aspects of the physical world that are relevant to its current task. Some of this knowledge will be derived from the robot's sensors, some of it perhaps directly entered into the environment model by the robot's designers, and some of it (as we will see) obtained from other humans through the mediation of the dialogue model. This model is expressed in an object-oriented formalism; the knowledge contained in the other two models is procedural, and is expressed in a Rule Description Language (RDL).

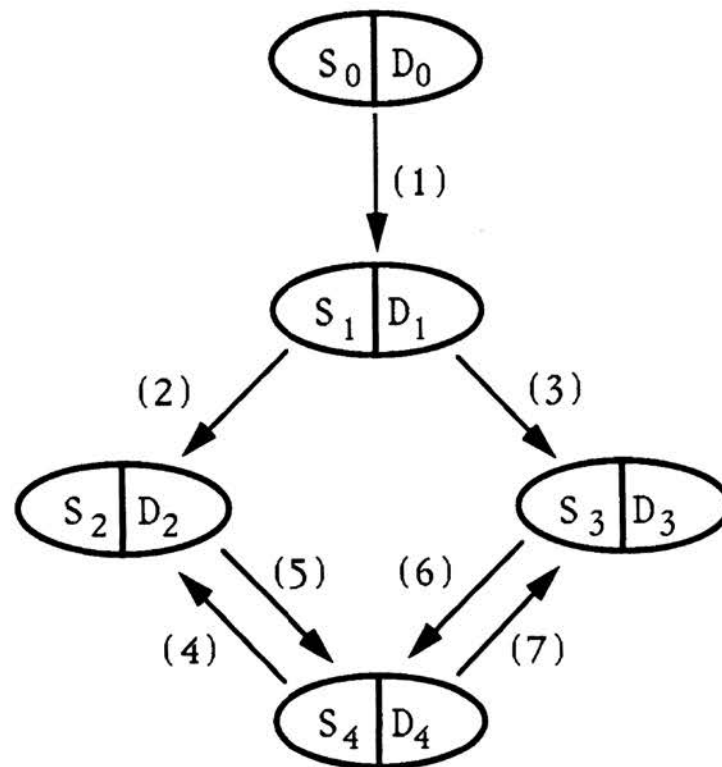
The **robot state model** contains the system's knowledge about what the robot can do and what it is doing now. In other words, it includes information about the states the robot can potentially be in, the nature of the transitions between states, and the robot's current state. It can be represented by an Augmented Transition Network Diagram that corresponds to a set of rules in the RDL.

The accompanying diagram (Figure 1) shows an example of an ATND. Most of the states S_i are associated with a Dialog Segment Manager D_i . Transitions between states are triggered by the setting of associated conditions to "true"; a condition may be set to "true" by a DSM (as in the example), or because of an observation made by the robot about its environment. In the example, state S_0 is the waiting state. The robot undergoes a transition to state S_1 when the user indicates that he wants the robot to undertake a task. In D_1 , associated with S_1 , the robot obtains a general description of this task from the user. In the example, there are two types of tasks: "Navigation" and "Manipulation". The corresponding states, S_2 and S_3 , are each associated with DSMs that obtain more precise definitions of the task from the user.

In this schema, each Dialog Segment Manager D_i contains the system's knowledge about how to carry out a dialogue of a particular kind. This knowledge has two components: knowledge about what the user is likely to say next, and knowledge about how to fulfill the robot's dialogue goals by asking questions or providing the user with needed information. The former consists of a set of conceptual expectations, C_i . Each element of C_i is a different hypothesis about the semantic content of what the user is about to say; each conceptual hypothesis has an a priori probability. The conceptual hypotheses are implemented as frames. Each is associated with a **Conceptual Hypothesis Grammar** that will work in association with the speech recognition component to change the probability of that hypothesis and fill in its slots. Thus, as the user speaks, certain conceptual hypotheses which seem to match the input well are assigned higher probabilities, and their slots are filled in.

The different DSMs employ similar conversational strategies for extracting information from their human interlocutors in order to fill in slots for these active conceptual hypotheses, and to determine which was intended by the user. When the Dialog Segment terminates, the DSM will usually pass only the best scored conceptual hypothesis to working memory (unless the DS was terminated prematurely and will be resumed later).

Figure 1.

Key

c = condition, a = action, o = operation

Arcs

- (1) c1 = recognized intention to define a task
- (2) c2 = user has given NAVIGATION goal
o = push NAVIGATION
- (3) c3 = user has given MANIPULATION goal
o = push MANIPULATION
- (4) c = user resumes or initiates NAVIGATION goal
o = push NAVIGATION
- (5) c = NAVIGATION complete or user interrupt occurs
- (6) c = MANIPULATION complete or user interrupt occurs
- (7) c = user resumes or initiates MANIPULATION goal
o = push MANIPULATION

Figure 2. Dialog Segment D_0

```

(start

  c1 := FALSE
  CHG := {Yes_no_answer}
  sentence := "Would you like me to carry out a task?"
  prepare_sentence(sentence)

  repeat

    output_sentence(sentence)
    activate(CHG)
    reply := speech_recognize(CHG)
    dialog_interpreter(reply,sentence,CHG,c1)
    prepare_sentence(sentence)

  until c1 = TRUE

  new_robot_status(S1)
)

```

Figure 3. Dialog Segment D_1

```

(start

  c2 := FALSE
  c3 := FALSE
  CHG := {Navigation_goal,Manipulation_goal}
  sentence := "What kind of task - navigation or manipulation?"
  prepare_sentence(sentence)

  repeat

    output_sentence(sentence)
    activate(CHG)
    reply := speech_recognize(CHG)
    dialog_interpreter(reply,sentence,CHG,c1,c2)
    prepare_sentence(sentence)

  until ((c1 = TRUE) or (c2 = TRUE))

  if (c1 = TRUE) then new_robot_status(S2)
  else new_robot_status(S3)
)

```

Figure 2 shows the high-level code for D_0 (the DS associated with the resting state S_0 shown in Figure 1) as it will appear in the RDL. Similarly, Figure 3 gives the high-level code for D_1 .

In both cases, "speech-recognize" assesses the user's spoken input with the aid of the information in CHG and returns "reply", which is usually a single conceptual hypothesis, but may instead be a list of CHs, along with their probabilities. The procedure "dialog-interpreter" analyses "reply" on the basis of the current goals. If it decides the information is incomplete, it may formulate further speech output. On the other hand, if there is enough information, it will set a condition flag to "TRUE", so that the system undergoes transition to a new state.

In addition to local DSMs of the type just discussed, there is a global Dialog Manager which handles interrupts and resumes. An interrupt might occur because the user decides to change the topic, or because of an event in the environment. In either case, a new DS may be entered. The global DM knows which frames needed to complete a task have been only partially filled out; it ensures that these are resumed when the current DS is over.

It should now be clear how the dialogue model relates to the environment model and the robot state model. As the robot undergoes state transitions, various actions are triggered. Some of these are physical actions involving movement in space; others, triggered by new information provided by the robot's sensors, are updates of the environment model. A third kind of action employs the dialogue model.

This kind of action occurs when, for instance, the robot asks the human for further commands, or explains an obstacle it has encountered and asks for guidance. It may also occur in response to an initiative by the user. In each case, a particular DSM will be invoked. Which one is chosen depends to a large extent on the goals the robot is currently trying to satisfy. These goals involve information that the robot needs to carry out its physical tasks, and can be broken into two categories: either the robot needs to be given a command, or it has been blocked in the performance of a task by inadequate information about the environment. When its goals have been satisfied, the robot will try to terminate the Dialog Segment unless the human provides indications that he or she has relevant information to add.

A successful dialogue may lead to a robot state transition (the robot is starting a new task), to an update of the environment model, or both. Note that in the course of a dialogue, the system is highly goal-oriented: it does its best to extract needed information from the user, asking as many questions as necessary to do so. This feature of our dialogue model is shared with many natural language systems, especially natural language systems with a speech recognition component. If the user changes topic, the system will deal with the new topic, but it will eventually revert to the old topic as long as it still has unsatisfied conversational goals that have not been rendered irrelevant by new information.

5 The Conceptual Hypothesis Grammar

The CHG associated with each conceptual hypothesis is a crucial part of the dialogue model. It has three main functions: to provide the speech recognition component with constraints on its search through the space of possible words, to evaluate the probability score of the hypothesis based on information provided by the speech recognition component, and to fill in slots in the hypothesis if this probability score is sufficiently high.

The CHG is a two-level stochastic grammar. Hypothesis H may be realised in a number of different ways at the grammatical level. For instance, consider the number of different ways of saying "Yes": "Certainly", "I would like that very much", "I guess so", and so on. These realisations differ widely in grammatical structure; we will name them $G_1, G_2, \dots, G_n$.

A particular grammatical structure G_i can in turn be parsed into a number of constituents:

$$G_i = \langle g_{1i}, g_{2i}, \dots, g_{ki} \rangle. \quad (1)$$

Finally, each one of these constituents g_{ji} yields a terminal. A terminal in a CHG could be a single word or a string of words, since certain multiword phrases in English recur so often that it makes sense to treat them as vocabulary items with an associated grammatical category. We will use the word "phrase" and the symbol p to refer to a terminal, with the understanding that this could represent either a single word or a string of words. The category g might yield a number of different terminals p . For instance, there may be a category "request-start" which includes such phrases as "please", "would you", "what I want is for you to", etc.

The CHG for a given hypothesis H assigns a probability $P(G_i|H)$ to each grammatical structure G_i . Now, suppose that the speech recognition component is evaluating the hypothesis that the speaker's last utterance was

$$U = \langle p_1, p_2, \dots, p_k \rangle. \quad (2)$$

Assume that p_1 is of grammatical category g_1 , p_2 is of category g_2 , and so on, and that this sequence of g 's has the structure

$$G = \langle g_1, g_2, \dots, g_k \rangle. \quad (3)$$

Clearly, we have

$$P(U|H) = P(G|H)P(p_1|g_1)P(p_2|g_2)\dots P(p_k|g_k) \quad (4)$$

Now, consider the probabilities $P(p_i|g_i)$. We can divide these probabilities into two classes: those that depend on the circumstances of the dialogue (such as dialogue history and the environment) and those that do not. The first class contains slot-filling phrases and the second contains all other phrases, with some exceptions (these will be phrases that do not fill slots but are still influenced by dialogue history). For instance, the choice of p_1 , p_3 , and p_6 might be dialogue-dependent. Let $P_{dep} = \prod_i P(p_i|g_i)$ for all such dialogue-dependent p_i s and g_i s, and let $P_{ind} = \prod_j P(p_j|g_j)$ for the rest. Then we have

$$P(U|H) = P(G|H)P_{dep}P_{ind}. \quad (5)$$

Thus, P_{dep} is the component of the CHG that is most important for slot-filling. Consider a particular factor of this probability, $P(p_i|g_i)$. This gives the probability that, among type g_i phrases, phrase p_i will be uttered by the user. In our model, this probability is given as a weighted average of three components, $P_S(p_i|g_i)$, $P_K(p_i|g_i)$, and $P_C(p_i|g_i)$.

Here subscript S stands for "static", K stands for "knowledge", and C stands for "cache". The formula is:

$$P(p_i|g_i) = \lambda_1 P_S + \lambda_2 P_K + \lambda_3 P_C \quad (6)$$

where $\lambda_1 + \lambda_2 + \lambda_3 = 1$.

The static component is an a priori probability derived from the frequency of p_i among type g_i phrases in the past. The knowledge component reflects information about the robot's current situation, including relevant facts about the environment. For instance, if a noun phrase of type OBJECT is being considered, we would assign higher probabilities to phrases describing objects in the robot's vicinity. Finally, the cache component reflects phrase associations.

The idea underlying this component is that the occurrence of a given phrase may be used to predict the occurrence of other phrases; it can be formalized as follows. Suppose that the Viterbi algorithm or some other probabilistic algorithm has identified the most likely interpretation of the acoustic signals up to now as the sequence $p_1, p_2, \dots, p_n$, where each phrase p_i is associated with a probability $P(p_i)$ derived from both acoustic and higher-level evidence. We are trying to identify phrase p_{n+1} . We consider only the L -sized window given by $p_{n-L+1}, p_{n-L+2}, \dots, p_n$, discarding earlier information. Within this window, some phrases are more likely to have been correctly identified than others; $P(p_i)$ indicates the strength of the evidence that p_i was correctly identified.

Now, define $Q(p, p_j)$ as the strength of the likelihood that phrase p will occur, given that phrase p_j occurred among the last L phrases. This can be calculated from a training text by setting up a counter matrix $c_x(p_i, p_j)$ for each DS x , which is incremented every time the phrase p_i occurs while p_j is contained in the last L phrases in the training text. Q for DS x can be defined as

$$Q_x(p_i, p_j) = c_x(p_i, p_j) / (\sum_i c_x(p_i, p_j)). \quad (7)$$

Finally, we define a function

$$f_L(p, < p_{n-L+1}, \dots, p_n >) = \sum_{j=1}^L P(p_{n-L+j}) Q(p, p_{n-L+j}). \quad (8)$$

This function (which requires a normalization constant not given here to become a true probability) uses each of the previously recognized phrases within the L -sized window to estimate the probability that phrase p will occur; it assigns higher weight to evidence from those previously recognized phrases considered to have a high probability of being correct.

Thus, the occurrence of a phrase carries with it increased probabilities for other phrases which have often followed it within a window of size L while a certain topic is being talked about by the user. Even if a few phrases are inserted or deleted, the statistical association between two phrases yields information (within the limit imposed by the window size L). For example, suppose L is 5 and the training corpus includes several occurrences of strings where phrase p_a is followed by a gap of one or two other phrases followed by p_b . Suppose that during recognition, p_a is recognized with high estimated probability; p_b is then assigned high probability of occurring after p_a with a gap of one or two other phrases, but in addition it is assigned high probability of occurring with a zero gap, a gap of three, and a gap of four.

The Dialog Segment Manager that employs the CHG interacts directly with the speech recognition component, each guiding the other in its search through the space of hypotheses. The speech recognition component provides the DSM with a list of the currently hypothesized words, each accompanied by its corresponding acoustic probability. In turn, the DSM activates a list of words in the vocabulary considered most likely to follow the current word on the basis of the CHG. When the user's utterance has terminated, the DSM may retain several alternative CHs whose probability is sufficiently high to have survived the selection process. In that case, the DSM will probably ask further questions to eliminate all but one of the CHs.

In theory, the probability of a CH should be the sum of the probabilities of all parsings of the user utterance that yield that CH. For reasons of computational efficiency, it seems likely that most low-probability parsings will never be carried out, so the score associated with a CH will be the probability estimate of the highest-probability parsing with that interpretation.

6 Conclusions

The difficulties of the task faced by a system for spoken man-machine dialogue are well-known. They include not only the existence of multiple word candidates at each location in time, mentioned above, but repetition of syllables or of whole words, deletion of syllables or of whole words, and insertion of interjections like "ah" or "um" or of words or phrases not recognised by the system. The system must therefore be robust enough to generate a semantic token even when semantically unimportant parts of the phrase have been distorted by repetition, deletion, or insertion.

As the research surveyed in the first two sections indicates, this can be achieved only if there are strong constraints on the space of hypotheses examined by the speech recognition component. The model described above contains a hierarchy of constraints, each given probabilistically. At the uppermost level, the system uses its knowledge of dialogue structure to guess what topic the user is about to address; furthermore, it seeks to guide the user's responses through its own voice output (without being excessively domineering). Given a topic, the system can estimate how the user is likely to structure his response. Finally, at the phrase level, the system uses three kinds of constraints: an a priority probability distribution for phrases derived from past phrase frequencies, a probability distribution based on word association frequencies, and a probability distribution based on the current situation of the robot.

This model reflects some of the key features of natural language. It is clear that the use of language by human beings is very dependent on context: on the physical environment, on the recent experiences of the speaker, on the mood of the speaker, on what has been said by the speaker's interlocutors, and so on. One could sum up the effects of all these factors by saying that they determine what is "salient" to the speaker and the listeners at a given time and place. This is one of the most important problems in computational linguistics, and it manifests itself at many different levels - for instance, anaphor is one aspect of it. The model described above should give the system a method for estimating when certain words or phrases are likely to be spoken - in other words, when they are salient.

Thus, the dialogue model in general and the estimated probabilities for words and phrases in particular will respond as dynamically as possible to the recent history of the robot and to its physical environment. For instance, when the robot's vision system observes a physical obstacle blocking its path, the associated words and concepts will be assigned higher probabilities. In addition, the user's most recent use of words and phrases will be taken into account.

The model requires the assignment of a large number of probabilities. As far as possible, these probabilities will be obtained empirically from studies of transcripts of robot-user sessions. Thus design of the system will be iterative - as experience with the system grows, the probability estimates should get more and more accurate.

We have described a model for the incorporation of semantic and pragmatic information in a speech recognition system via explicitly specified probabilities. Implementation of the model will, we hope, yield performance improvements similar to those obtained by probabilistic approaches at other levels of the speech recognition hierarchy. As the use of voice-activated devices spreads to non-expert users, this kind of model is likely to receive more and more attention.

7 Acknowledgements

Acknowledgements to Charles Snow for attentive proofreading, to Diane Goupil for helpful comments, and to Henriette Bouchard and Sylvie Tellier for editing the bibliography.

8 Bibliography

1. Richard A. Bolt. "The integrated multi-modal interface" in *The Transactions of the Institute of Electronics, Information and Communication Engineering*, Vol. J70-D, no 11, pp. 2017-2025 (1987)
2. Harry Bunt. "On-line Interpretation in Speech Understanding and Dialogue Systems" in *Recent Advances in Speech Understanding and Dialog Systems*, NATO ASI Series, Series F; Computer and Systems Sciences, Vol. 46, pp. 348-395 Springer-Verlag (1988)
3. H.C. Bunt. "Information Dialogues as Communicative Action in Relation to Partner Modeling and Information Processing" in *The Structure of Multimodal Dialogue*, M.M. Taylor, F. Néel and D.G. Bouwhuis (Editors) Elsevier Science Publishers B.V. (1989)
4. N. Carbonell and J.M. Pierrel. "Task-oriented dialogue processing in human-computer voice communication" in *Recent Advances in Speech Understanding and Dialog Systems*, NATO ASI Series, Series F; Computer and Systems Sciences, Vol. 46, pp. 491-496 Springer-Verlag (1988)
5. N. Carbonell and J.M. Pierrel. "Architecture and Knowledge Sources in a Human-Computer Oral Dialogue System" in *The Structure of Multimodal Dialogue*, pp. 417-429, M.M. Taylor, F. Néel and D.G. Bouwhuis (Editors) Elsevier Science Publishers B.V. (1989)

6. Guy Deville. "Modélisation d'énoncés finalisés dans un système de dialogue oral homme-machine" in XVIIIèmes Journées d'études sur la parole, pp. 242-247 (Conference held at Université de Montréal, May 28-31, 1990)
7. Jack L. Edwards and James A. Mason. "Toward intelligent dialogue with ISIS" in International Journal of Man-Machine Studies, 28, pp. 309-342 Academic Press (1988)
8. J.L. Edwards and J.A. Mason. "The Structure of Intelligence in Dialogue" in The Structure of Multimodal Dialogue, M.M. Taylor, F. Néel and D.G. Bouwhuis (Editors) Elsevier Science Publishers B.V. (1989)
9. U. Ehrlich and H. Niemann. "Using Semantic and Pragmatic Knowledge for the Interpretation of Syntactic Constituents" in Recent Advances in Speech Understanding and Dialog Systems, NATO ASI Series, Series F; Computer and Systems Sciences, Vol. 46, pp. 485-490 Springer-Verlag (1988)
10. P. Falzon. "Ergonomie cognitive du dialogue". 175 p. Presses Universitaires de Grenoble (1989)
11. M. Guyomard and J. Siroux. "Experimentation in the specification of an oral dialogue" in Recent Advances in Speech Understanding and Dialog Systems, NATO ASI Series, Series F; Computer and Systems Sciences, Vol. 46, pp. 497-501 Springer-Verlag (1988)
12. M. Guyomard, J. Siroux and A. Cozannet. "Le rôle du dialogue pour la reconnaissance de parole: Le cas du système pages jaunes" in XVIIIèmes Journées d'études sur la parole, pp.322-326 (Conference held at Université de Montréal, May 28-31, 1990)
13. J. Klein and J.M. Pierrel. "Interprétation d'expressions complexes dans un système de dialogue homme-machine" in XVIIIèmes Journées d'études sur la parole, pp. 248-252 (Conference held at Université de Montréal, May 28-31, 1990)
14. Roland Kuhn and Renato De Mori. "A Cache-Based Natural Language Model for Speech Recognition" in IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 12, no 6, pp. 570-583 (June 1990)
15. S. Kunzmann, T. Kuhn and H. Niemann. "An experimental environment for the generation and verification of word hypotheses in continuous speech" in Speech Communication 7, pp. 381-388 North-Holland (1988)
16. Akira Kurematsu. "An overview of ATR basic research into telephone interpretation", Keynote Speech in ATR Symposium on basic research for telephone interpretation, Kyoto, Japan, December 11-12, 1989, 14 p.
17. James A. Mason and Jack L. Edwards. "Surveying projects on intelligent dialogue" in International Journal of Man Machine Studies, 28, pp. 259-307 Academic Press (1988)

18. K. Matrouf et al. "Adapting probability-transitions in DP matching process for an oral task-oriented dialogue" in Proceedings of ICASSP 90. New- Jersey : IEEE, 1990 : pp. 569-571 (Conference held at: New Mexico, April 3-6, 1990)
19. Karim Matrouf, Françoise Néel and Joseph Mariani. "Système de dialogue orienté par la tâche: une application en aéronautique" in Journal Accoustique 2, pp. 85-93 (mars 1989)
20. Robert Moore, Fernando Pereira and Hy Murveit. "Integrating Speech and Natural-Language Processing" in Proceedings of the Speech and Natural Language Workshop. California : Morgan Kaufmann, pp. 243-247. Sponsored by DARPA/ISTO. (Conference held at: Philadelphia, Pennsylvania, February 21-23, 1989)
21. Hy Murveit and Robert Moore. "Integrating Natural Language Constraints into HMM-based Speech Recognition" in Proceedings of ICASSP 90. New-Jersey : IEEE 1990 : pp. 573-576. (Conference held at: New Mexico, April 3- 6, 1990)
22. H. Niemann et al. "The Speech Understanding and Dialog System EVAR" in New Systems and Architectures for Automatic Speech Recognition and Synthesis, R. de Mori and C.Y. Suen (eds). NATO ASI Series F16, Berlin, pp. 271-302
23. J.M. Pierrel. "Dialogue oral Homme-Machine: connaissances linguistiques, stratégies et architectures des systèmes", Hermès, Paris (1987)
24. Salim Roukos. "Integrating Speech and Natural Language" in Proceedings of the Speech and Natural Language Workshop. California : Morgan Kaufmann, pp. 243-247. Sponsored by DARPA/ISTO. (Conference held at: Philadelphia, Pennsylvania, February 21-23, 1989)
25. Alexander I. Rudnicky. "The Design of Voice-Driven Interfaces" in Proceedings of the Speech and Natural Language Workshop. California : Morgan Kaufmann, pp. 120-124. Sponsored by DARPA/ISTO. (Conference held at: Philadelphia, Pennsylvania, February 21-23, 1989)
26. Alexander I. Rudnicky, Michelle Sakamoto and Joseph H. Polifroni. "Spoken Language Interaction in a Goal-Directed Task" in Proceedings of ICASSP 90. New-Jersey : IEEE, 1990 : pp. 45-48 (Conference held at: New Mexico, April 3-6, 1990)
27. G. Sagerer and F. Kummert. "Knowledge Based Systems for Speech Understanding" in Recent Advances in Speech Understanding and Dialog Systems, NATO ASI Series, Series F; Computer and Systems Sciences, Vol. 46, pp. 421-458 Springer-Verlag (1988)
28. Stephanie Seneff. "TINA: A probabilistic syntactic parser for speech understanding systems" in Proceedings of the Speech and Natural Language Workshop. California : Morgan Kaufmann, pp. 168-178. Sponsored by DARPA/ISTO. (Conference held at: Philadelphia, Pennsylvania, February 21-23, 1989)

29. Jacques Siroux and Dominique Gillet. "A system for man-machine communication using speech" in *Speech Communication*, Vol. 4, no 4, pp. 289-315 North-Holland (December 1985)
30. Gilles Souvay, Jean-Marie Pierrel, Evelyne Gallais and Pierre Alinat. "Diapason: un système de DIALogue Pour la commande orAle d'une console SONar" in *XVIIIèmes Journées d'études sur la parole*, pp. 327-331 (Conference held at Université de Montréal, May 28-31, 1990)
31. M.M. Taylor. "Layered protocols for computer-human dialogue. I: Principles" in *International Journal of Man-Machine Studies*, 28, pp. 175-218 Academic Press (1988)
32. M.M. Taylor. "Layered protocols for computer-human dialogue. II: some practical issues" in *International Journal of Man-Machine Studies*, 28, pp. 219-257 Academic Press (1988)
33. M.M. Taylor, F. Néel and D.G. Bouwhuis (Editors). "The Structure of Multimodal Dialogue" Elsevier Science Publishers B.V. (1989)
34. G. Thurmair. "Semantic Processing in Speech Understanding" in *Recent Advances in Speech Understanding and Dialog Systems*, NATO ASI Series, Series F; Computer and Systems Sciences, Vol. 46, pp.397-419 Springer- Verlag (1988)
35. Wolfgang Wahlster and Alfred Kobsa. "User Models in Dialog Systems", Technical Report No. 30, Universitat des Saarlandes, January 1988.
36. Wayne Ward. "Understanding Spontaneous Speech" in *Proceedings of the Speech and Natural Language Workshop*. California : Morgan Kaufmann, pp. 137-141. Sponsored by DARPA/ISTO. (Conference held at: Philadelphia, Pennsylvania, February 21-23, 1989)
37. J. Waterworth. "Interactive Strategies for Conversational Computer Systems" in *The Structure of Multimodal Dialogue*, M.M. Taylor et al. eds, North-Holland (1988)
38. S.J. Young and C.E. Proctor. "The design and implementation of dialogue control in voice operated database inquiry systems" in *Computer Speech and Language*, Vol.3, no 4 , pp. 329-353 Academic Press (October 1989)
39. Sheryl R. Young, Wayne H. Ward and Alexander G. Hauptmann. "Layering predictions: Flexible use of dialog expectation in speech recognition" in *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, Vol. 2 (Conference held in Detroit, Michigan, August 20-25, 1989)
40. Sheryl R. Young. "The Minds System: Using Context and Dialog to Enhance Speech Recognition" in *Proceedings of the Speech and Natural Language Workshop*. California : Morgan Kaufmann, pp. 131-136. Sponsored by DARPA/ISTO. (Conference held at: Philadelphia, Pennsylvania, February 21-23, 1989)

41. S.R. Young et al. "High Level Knowledge Sources in Usable Speech Recognition Systems" in Communications of the ACM, Vol. 32, no 2, (1989)
42. Victor Zue and al. "The Voyager Speech Understanding System: Preliminary Development and Evaluation" in Proceedings of ICASSP 90. New Jersey : IEEE, 1990 : pp. 73-76. (Conference held at: New Mexico, April 3-6, 1990)