Speech Understanding and Dialogue over the telephone: an overview of the ESPRIT SUNDIAL project.

Jeremy Peckham

Logica
Betjeman House
104 Hills Road
Cambridge CB2 1LQ

1. Introduction

One of the most obvious and natural applications for speech technology is in providing a gateway to information services over the telephone network. Already a significant growth is occurring the provision of information from a centralised computing system using stored messages or synthetic speech derived from text files. Most of these systems however currently rely on the use of touch tone input for selection of the information. In the UK only 15% of homes and businesses have tone dialling, although the network is currently able to support around 50% tone dialling. The ability to recognise a small number of words or even the digits without the user requiring to train the system therefore has widespread application. As the scope of the information service expands so also does the need for more intelligent dialogues with much larger vocabularies for speech understanding.

The SUNDIAL project, led by Logica Cambridge, is currently one of Europe's largest collaborative projects in speech technology involving partners from the UK, France, Germany, Italy and Sweden in 170 man years of effort over five years. The project which is part funded by the Commission of the European Communities under the ESPRIT programme, commenced in September 1988 and will end in July 1993. The partners involved in the project are shown below.

• UK	Logica (Project Leader)
	University of Surrey
• France	CNET
	CAP SESA Innovation
	IRISA - University of Rennes
• Italy	CSELT
	Saritel
	Politechnico di Torino (associate partner)
• Germany	Daimler Benz
	Siemens
	University of Erlangen
• Sweden	Infovox (subcontractor)

The computer systems which the SUNDIAL research aims to produce will enable users to maintain telephone conversations about specific topics such as flight arrivals, schedules and

reservations. The systems are planned to support a speaker independent vocabulary of up to 1000 - 2000 words in four languages (English, French, German and Italian).

The project builds on experience gained by partners in previous ESPRIT projects such as P26 - "Advanced Algorithms and Architectures for Speech and Image Processing", P316 - ESTEAM and P1015 - "Palabre", as well as other projects funded by national programmes such as VODIS (UK) and SPICOS (FRG) (Fissore et al 88, Niemann et al 88, Peckham 89, Brenner 89) A number of these projects have already demonstrated speaker dependent speech understanding with vocabularies of around 1000 words.

2. Architecture

A common architecture has been agreed amongst the partners in order to fulfil two main purposes;

- to give the detailed specifications of the interfaces between the major modules (front end, linguistic processing, dialogue management and message generation) thus allowing exchanges and comparative assessment.
- to allow the implementation of different processing strategies by modifying the control parameters and the flow of data between the modules.

The architecture is of a 'Distributed Database ' type which uses bi-directional interfaces, currently only between each major module. The Distributed Database Architecture (DDA) is most fully exploited at the present time within the Dialogue Manager module and this is described more fully in a section 6.

The major modules and architecture of the system are shown in figure 1. These modules make use of knowledge-based techniques for dialogue management and rule-based as well as statistical natural language processing. Recognition is based on an acoustic pattern-matching technique which uses Hidden Markov Models (HMM's) of phoneme sized speech units, thus providing a means of more efficiently handling a large vocabulary and the coarticulation effects associated with fluent speech. It is planned to carry out experiments during 1991 on the use of predictions at various levels within the architecture. These include passing down predictions generated by the dialogue module to the Linguistic processor. This module may in turn provide lexical constraints at the acoustic-word decoding stage.

Four separate language demonstrators are being built with applications covering Intercity Train times (Germany and Italy) and Flight Enquiries/Reservations (UK &France). The acoustic-phonetic decoding is aimed at speaker independence with the aim of operation over the telephone network. In all cases the goal is to recognise naturally spoken sentences, relevant to the chosen application domains. The interactive dialogue component is expected be a crucial part of any robust speech based computer information system, providing graceful error recovery and dialogue repair as well as eliciting further information, handling clarification and confirmation.

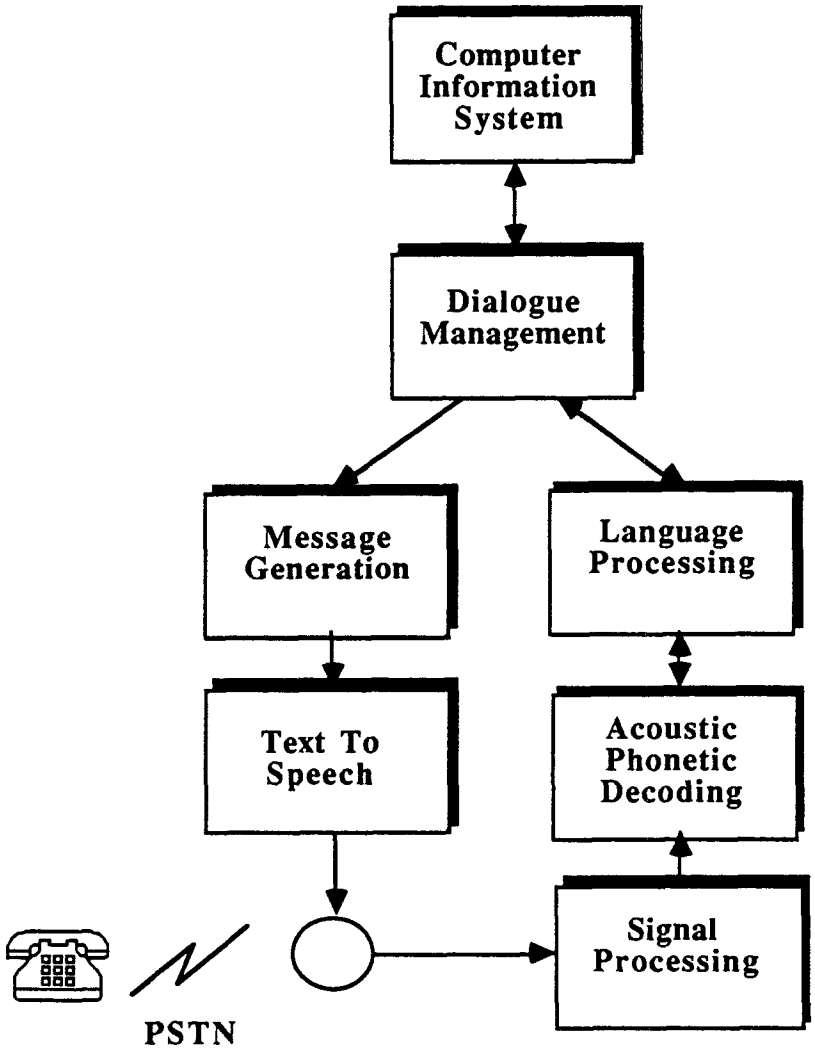

figure 1: - SUNDIAL Architecture

Examples of the type of dialogue which SUNDIAL aims to support are shown below for a flight enquiries application.

AGENT: Flight information.

USER: er:m I'd like some information on flight number AJ296 er it departs from Chicago and arrives in London today.
er:m I would like the exact arrival time please.

AGENT: Flight AJ296 from Chicago to London Heathrow terminal four is running ahead of schedule it will be arriving at ten fifteen.

USER: Is that one fifteen er AM or PM

AGENT: ten fifteen AM

3. Application studies and simulation

One of the key motivations of the project is to base the technology research and development on a clear understanding of user requirements and the spoken language phenomena for each of the selected application domains. In order to provide insight into user requirements a number of 'Wizard of Oz' simulations have been performed for some of the selected applications. In addition analysis has been carried out on a large corpus of human-human dialogues relating to flight enquiries and reservations, in order to inform the simulations and also to provide a baseline for unconstrained dialogues.

The methodology for Wizard of Oz simulations is that a human accomplice takes the role of the computer speech understanding component and the output to the user is provided by either synthetic speech or vocoded natural speech (Fraser and Gilbert, 90). Subjects are presented with various scenarios to enable them to make enquiries about flights or hotels. The human accomplice is given certain constraints to attempt to emulate the performance of a speech understanding system. These constraints, in some of the simulations, were imposed through a menu-driven tool which restricted the flexibility of the systems output at any stage in the dialogue. (Ponamale et al, 90). A small database was also used to access the requested information.

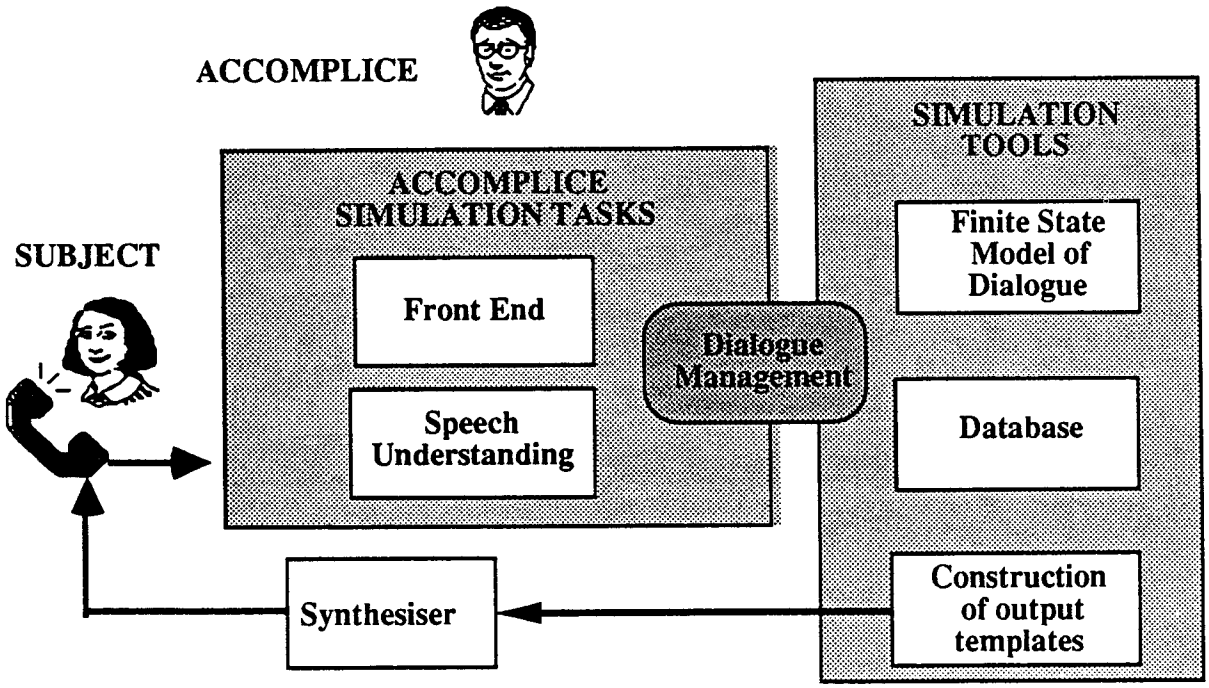

Figure 2: - Wizard of Oz Simulation Scenario

The simulations therefore provide a useful means of assessing the potential impact on users of the expected limitations or constraints of a future computer speech understanding and co-operative dialogue system. The corpora derived from these simulations have been analysed to define the spoken language phenomena which occur, as well as the required lexicon, grammar, semantics and dialogue rules for the chosen applications.

It is already clear from analysing the corpora from the first round of simulations that many of the requirements in terms of dialogue strategies are common to the four languages studied and to the different applications. It is expected that these requirements will also prove to be generic to other information service domains.

4. Speech pattern processing

The Front End Processing (FEP) module carries out the acoustic-phonetic decoding of the incoming speech signal and produces a lattice or graph of word hypotheses. Techniques for handling the variable quality of telephone speech, mostly due to the use of different handsets and varying handset positions between calls, are also included in this module. The FEP for all four languages is based on Hidden Markov Modelling (HMM) of sub-word units, usually context sensitive phone models. Some special cases such as function words and digits are modelled separately to improve performance.

Particular objectives in the speech pattern processing module include:

- refinement of the acoustic/phonetic units used, in line with requirements for easy extension of the vocabulary, rapid speaker adaptation and handling fluent speech (the co-articulation problem).
- improvement of speaker-independent phonetic rules and classification techniques.
- improvements in speech recognition over the telephone.
- fast lexical access.
- improvements in speech pattern processing techniques based on Hidden Markov Modelling including comparison and trade offs between Continuous Density HMM (CDHMM) and Discrete Density HMM (DDHMM) techniques and Vector Quantisation.
- real-time performance on DSP and/or Transputers.

A number of experiments have been divided amongst partners to address these objectives where the issue is expected to be language independent (Table 1).

DDHMM vs CDHMM vs Semi CDHMM
Vector Quantisation (Soft VQ, VQ of distributions)
Model Topology
No. of mixtures for CDHMM
Variable Frame Rate Analysis
Recognition algorithms (eg N best word chains)
Linguistic constraints (including statistical grammars)
Training Algorithms and Methodlogy

Table 1: - HMM Experiments

Language dependent issues, such as the precise inventory of speech units to be used, are being investigated independently for each language. Some experimentation is being done with 'context independent' phone models using mixture distributions for the CDHMM and training the phones in many different contexts. A typical training procedure is shown in figure 3.

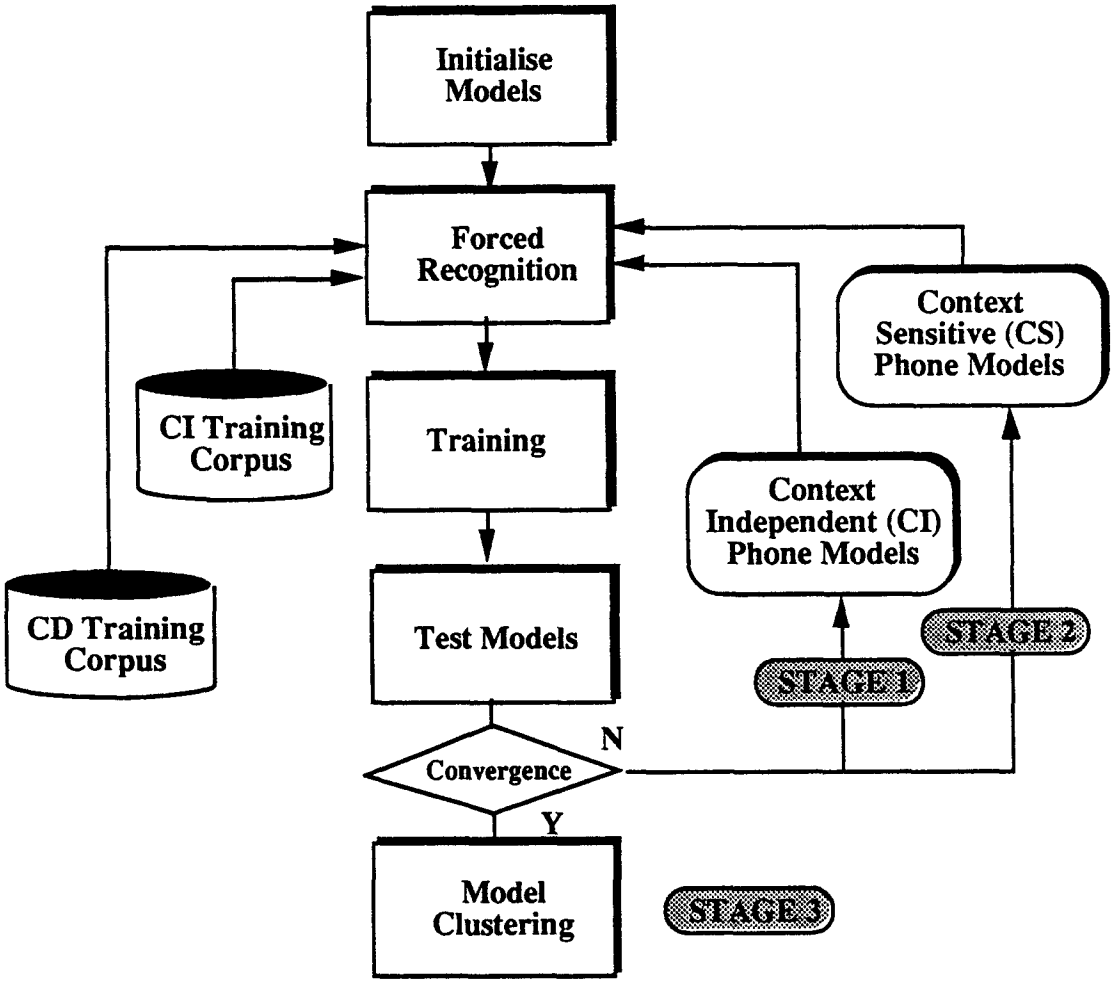

Figure 3: - A typical HMM Training Procedure

To support the development of speaker independent phone models a number of large multispeaker corpora have already been recorded for each language, mostly over the telephone. These contain a set of phonetically balanced sentences for each language (for English these are similar to the TIMIT and SCRIBE sentences) and a set of application specific sentences.

Preliminary results for the acoustic-phonetic decoding module show that continuous density HMM's (CDHMM) achieve 75.83% word accuracy on sentences, compared to 67.13% for discrete density HMM's, using 305 phonetic units for the Italian language and around a 1000 word vocabulary (Fissore et al, 90). Word accuracy includes insertions, deletions and substitutions. These results are for speaker independent recognition of telephone quality sentences; however they do not take into account the effect of the linguistic processing module on sentence understanding performance.

Some benchmarking experiments have been done and is also ongoing using the DARPA TIMIT and Resource Management Databases. Results for the English language using CDHHM's show that phoneme recognition accuracy on the DARPA TIMIT database is comparable to that achieved by Kai-Fu Lee in the Carnegie Mellon SPHINX system.

The lexical access stage and the use of phonological knowledge in this process to allow pronunciation differences and co-articulation effects is shown in figure 4. A word graph resulting from this process is illustrated in figure 5. A common formalism has been agreed amongst all partners for the representation of this data of the form;

$$[(B,E)\ (string)\ (t_S,\ t_e)\ (score)]$$

where:

t_S, t_e	are the start frame and end frame respectively
B and E	are equal to t_S and t_e in the case of a linear word chain or are equal to the start and end nodes for a word graph
string	is a pointer to to a single word or sequence of words
score	is the acoustic likelihood or score

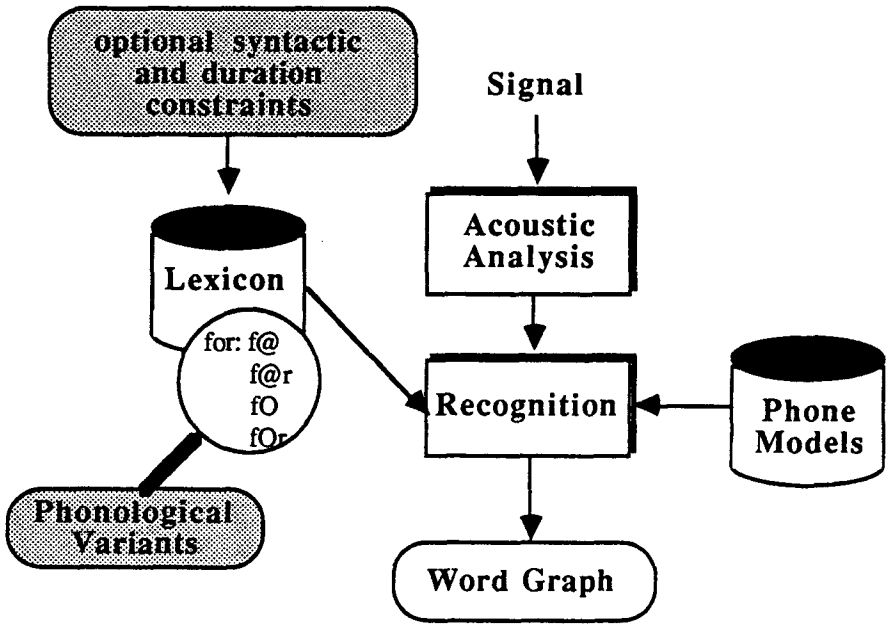

figure 4: - Lexical access

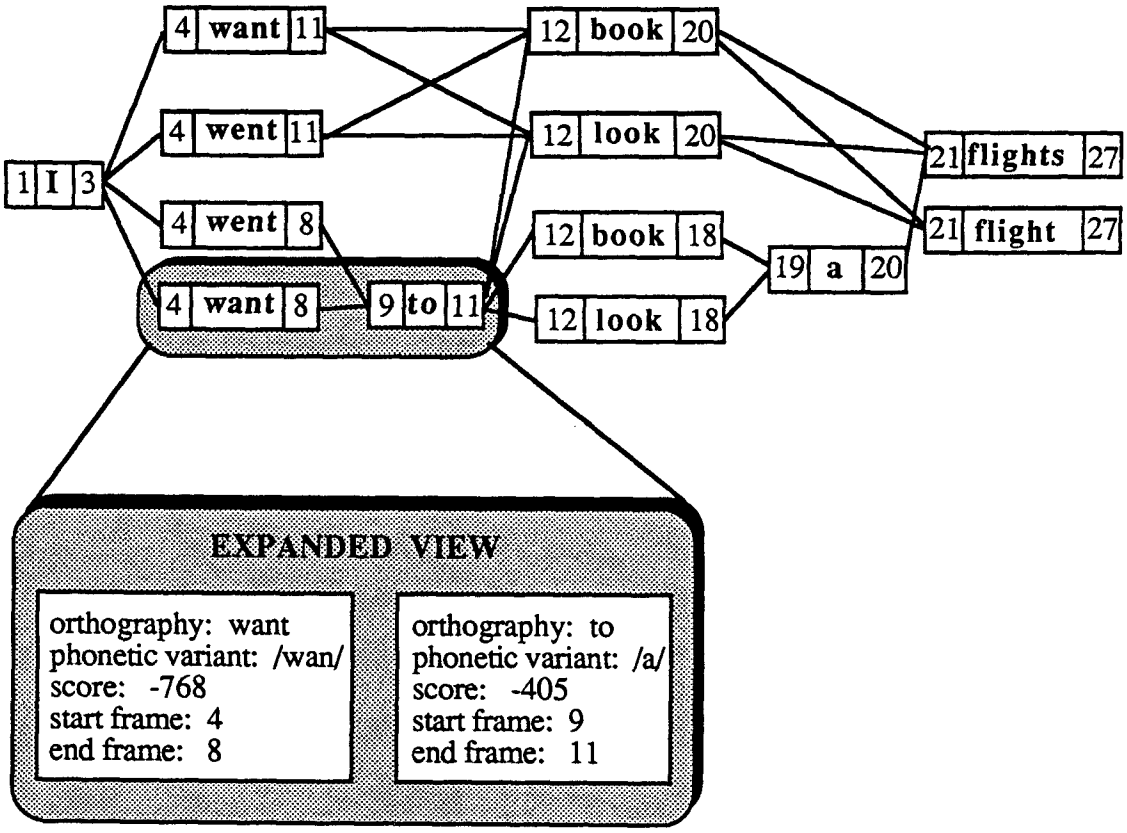

figure 5: - An example of a partial word graph

In co-operation with the ESPRIT Adverse Environment Recognition of Speech (ARS) project, an extensive evaluation of a number of acoustic analysis techniques has been carried out. This has used the Recogniser Sensitivity Analysis technique developed by Logica in the Alvey Speech Technology Assessment project (Peckham,90). These experiments were performed with a standard CDHMM modeler and recogniser using 6 mixtures and 8 states per word for a speaker independent isolated word recognition task. Tests on the recogniser using RSA have shown a best performance of 95.6% correct recognition ($\pm$ 0.7% at the 95% confidence level) on the RSA 31 word vocabulary. All tests were carried out with telephone quality speech.

Following this work a small vocabulary over-the-telephone speaker-independent recogniser has been developed with a vocabulary of around 50 words, suitable for a telephone banking application. A full continuous speech telephone banking application for English, with a vocabulary of around 300 words will be demonstrated at the next project review in October. This prototype will combine work on linguistic processing, dialogue management and message generation.

5. Linguistic processing

Considerable work has been carried out over the last 5 years within the ESPRIT programme as well as national programmes in the area of natural written language processing. This has resulted in a number of significant formalisms and grammar rules

for various European languages. Many of the SUNDIAL partners brought significant background into the project both in parsing technology and grammars and it was therefore uneconomic to consider a single formalism for the whole project. The principal formalisms currently in use in SUNDIAL are Unification Categorial Grammar - UCG (English and French), Augmented Phrase Structure Grammar - APSG (German) and Dependency Grammar - DG (Italian) (Poesio 87, Tropf 89, Zeevat 87). In the case of UCG the semantics are compiled into the lexicon whereas in both APSG and DG the semantics are represented separately as either a semantic net or case frames.

The major objectives of the linguistic processing module are:

- the modification of linguistic processing algorithms to meet the characteristics of spoken language such as hesitations and ill-formed or incomplete utterances. This may require, for example, the partial analysis of complete utterances;
- the development of analysis algorithms to handle indeterminacy within the linguistic analysis stage.
- the study and modelling of the spoken language sub-set used in the selected applications;
- exchange and evaluation of different parsing algorithms;
- development and evaluation of stochastic grammars in relation to rule based systems, with particular reference to computational efficiency, coverage, performance and extensibility;
- investigation and comparison of ways of integrating linguistic knowledge with pattern processing and with dialogue management;
- use of predictions from the linguistic processing module to improve hypotheses at the front end;

Two main classes of parser have been implemented: a left to right bottom up chart parser and an island driven parser which makes use of the best acoustic scoring hypotheses to determine starting points for parsing (Giachin 88). A fast 'C' version of the island driven parser has been implemented which can be ported to a transputer array. Typical parse times on a SUN 4 are on the order of a few seconds for a moderately complex sentence with a 1000 word lexicon. This version will be evaluated for both Italian and English. Experiments are also underway to explore the benefits of bigram statistics as an initial filter on word hypothesis generation.

6. Dialogue management

Human-human dialogue is capable of robust handling of adverse conditions and recovery from communication failure. In human-machine communication which is speech based, it becomes necessary to repair recognition failure (Young 89). Knowledge from a variety of sources: syntactic, semantic and pragmatic, and

knowledge of the application domain, may be brought to bear both in understanding and recovery.

A major goal of the project is the development of an intelligent co-operative dialogue system. Most of the work done to date in the area of dialogue management has been in text based systems. Whilst speech-based systems can learn much from this work, particularly in the areas of architectures and the use of dialogue histories, oral dialogue contains specific phenomena not seen in text, such as hesitations and false starts. Knowledge is also used differently, for instance pragmatic knowledge is not only used to solve ambiguities but also to strengthen or weaken hypotheses generated from potentially error-prone acoustic-phonetic decoding.

To progress in the modelling of oral dialogue, extensive use is being made of simulations and real applications in the selected information service domains such as flight enquiries, intercity train times and hotel information. A substantial amount of analysis is already complete on a large corpus of spoken dialogues for the four languages. This analysis is providing information on:

- user requirements;
- oral dialogue strategies; trade-offs between user and system initiative in terms of user acceptability and constraining the domain (necessary to avoid over ambitious questions), confirmation and repair strategies;
- oral language requirements (grammar, lexicon, specific phenomena of spoken dialogue);
- the complexity of the semantic space of the domains chosen;
- generic rules for dialogue for information services domains where these can be abstracted from particular applications, including cross-language differences in dialogue strategies.

It is a major goal of SUNDIAL to handle a number of the observed attributes of normal human-human dialogue such as: turn taking, anaphora and ellipsis (utterances which depend on their context for their meaning), hesitations, coughs, changing the subject, strategies for dealing with communication failure, implicit and explicit confirmation, and resolution of ambiguity via questioning. After some initial prototyping and the development of a common (to all partners) functional specification, a prototype Dialogue Manager which is based on a distributed database architecture has been implemented. In this approach, various modules or agents (shown in figure 6) communicate with each other, maintaining their own histories or knowledge bases (the so called distributed database).

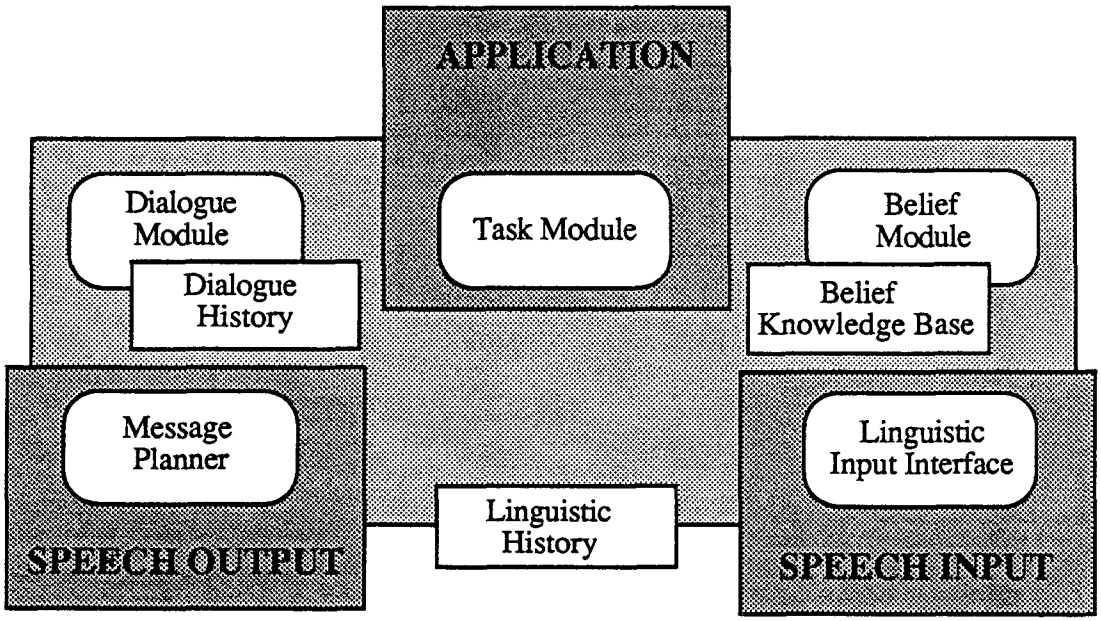

Figure 6: - Dialogue Manager Distributed Database Architecture

The system operates by interleaving control between agents; these communicate by message passing. In the current phase of implementation on serial hardware, only one agent is active at any one time.

Typically, the cycle is initiated by the task module, which makes a request for task information to the dialogue module. This module keeps a structured history of the dialogue so far, and is able to reason about possible next moves. It may propose a next move for the system and inform the message planner, and send a set of possible next caller moves to the linguistic interface module, as predictions. These predictions are elaborated into linguistic patterns that can serve to constrain the parser.

After the caller's input has been processed, it is assumed that a best prediction instantiation has been found, and this is translated into dialogic terms by the linguistic interface and passed back to the dialogue module. The semantic component of the recognised utterance is passed to the belief module, for updating of the belief state and resolution of anaphora and ellipses. Finally, the contextually elaborated message is passed on to the task module, which may need to go through several cycles of parameter acquisition and negotiation, before a database access can be made which satisfies the user's requirements.

7. Oral message generation

Whilst research on natural language generation is more recent than natural language parsing and understanding, computer based language generators are rapidly expanding in response to growing needs for intelligent human-machine interaction. The emphasis hitherto has again been on written language.

The particular requirements of oral output in the course of a dialogue (intelligibility, reduced length, enumeration and requests for repetition etc) is being taken into account

in the message generation component. Prosody (stress on words and melodic contour of an utterance) is affected by both linguistic and pragmatic constraints. A suitable symbolic description of the effects of these constraints on a particular utterance is passed, along with the text, to the Text to Speech Synthesis (TTS) system.(House and Youd 90).

The message generation module contains a message planning and message generation component. The message planner is responsible for formulating the output according to context and overall prosodic contour planning. In the future it will also handle issues such as summarising. The linguistic generator uses the 'pivot' generation algorithm for generation and also makes use of the same grammar formalism as the linguistic processing module. It handles ellipsis and refering expressions and also provides the final prosodically annotated output to the Text To Speech (TTS) synthesiser.

The text to speech synthesisers (for each of the four languages) are based on existing systems (diphone synthesis from CNET and formant synthesis from Infovox). Work is under way to improve the segmental quality of both the diphone and formant synthesis approaches and a preliminary comparative evaluation of quality using preference tests has already been carried out for the four languages. Much of the work for British English is being carried out by Infovox of Sweden who is a subcontractor to Logica.

8. Integration and real time constraints

A number of laboratory prototypes, in four languages are planned for July 1991 which will be designed to run in real-time or near-real-time principally for the purposes of evaluation. Progress on the co-operative dialogue element and its interaction with the lower modules can only realistically be evaluated with potential users. It is not intended however to develop special purpose hardware since this has already be done in a number of partners' laboratories. Specific hardware to be used includes transputer arrays, Digital Signal Processors (DSPs) and Sun workstations.

Careful consideration is being given to the requirements of interaction between symbolic and numeric processing (typically between the higher and lower levels, e.g. semantics, syntax and the acoustic pattern processing), the latter being performed on digital signal processing boards and transputers.

9. Conclusion

The SUNDIAL project is one of the most ambitious speech and language research projects in Europe, Japan or the USA. The US DARPA projects on spoken language understanding for example, have some similar goals to those of SUNDIAL (Zue 90, Price 90), however the major differences are that SUNDIAL places greater emphasis on the dialogue component and also attempts speech understanding over the telephone. Already encouraging results are being obtained within modules of the SUNDIAL system and excellent co-operation has been established between partners including exchange of software and techniques.

The first major milestone at which progress of complete systems will be demonstrated is in July 1991 At this point both flight enquiry and reservations as well as train timetable enquiry applications will be shown. These systems will form the basis for further refinement and development in the final phase of the project leading up to July 1993.

The pay-offs for success in achieving the goals of SUNDIAL could be considerable as telecommunications services expand in the 1990's. Despite the increasing take up of tone dialing in Europe and the popularity of terminal based services such as Minitel in France, public and professional access to information over standard telephones by speech is likely to find wide acceptance. The challenge is in matching the technology capability to the user requirements and market demand, to this end much work remains to be done.

References

Brenner, M. et al Word Recognition in Continuous speech using a phonological based Two-Network Matching Parser and a synthesis based Prediction, Proc. ICASSP, p457-460, 1989.

Fissore,L., Giachin,E., Laface,P., Micca,G., Pieraccini,R and Rullent,C Experimental results on large vocabulary continuous speech recognition and understanding, Proc. of the ICASSP '88, pp. 414-417, New York, NY April. 1988.

Fissore, L.,et al Performance of a Speaker Independent Continuous Speech Recogniser, Proc NATO ASI Speech Recognition and Understanding Recent Advances, Trends and Applications", Cetraro,Italy 1990.

Fraser, N, and Gilbert, G.N., Simulating Speech Systems, to appear in Computer Speech and Language.

Giachin,E.P., Rullent C., Robust Parsing of Severely Corrupted Spoken Utterances, in COLING, pp 196-201, Budapest, 1988.

House, J. and Youd, N. Contextually appropriate intonation in speech synthesis , Proc. ESCA: Tutorial and Workshop on Speech Synthesis, Sept 1990.

Niemann, H, Brietzmann, A, Ehrlich,U, Posch,S, Regal,P, Sagerer,G, Salzbrunn, R and Schukat-Talamazzini, G. A knowledge based speech understanding system, Int. J.. Pattern Recognition and Artificial Intelligence, 2(2): 321-350, 1988.

Peckham, J. VODIS - a Voice Operated Database Enquiry System, in Recent Developments and Applications of Natural Language Processing, ed. Jeremy Peckham. Kogan Page 1989.

Peckham, J.et al., Recogniser Sensitivity Analysis: A method for assessing the performance of speech recognisers, in publication - Speech Communication 1990.

Poesio,M., Rullent C., Modified caseframe parsing for speech understanding systems, Proc. IJCAI 87, Milano, 1987.

Ponamale M., Bilange,E., Choukri, K and Soudoplatoff, S A., Computer-aided Approach to the Design of an Oral Dialogue System, In Proceedings of Eastern Multi-conference 90: AI and Simulation, Nashville, Tennessee, pp229-232, 1990.

Price, P., SRI's Spoken Language System for Air Travel Planning,..Proceedings Speech Tech 90, 371-374, 1990.

Tropf, H.S., Syntax in the Spoken Dialogue System SPICOS-II, in Eurospeech 89, pp. 30-33, Paris, 1989.

Young, S.J., Proctor, C.E., The Design and Implementation of Dialogue Control in Voice Operated Database Inquiry Systems, Computer Speech and Language, 3,pp.329-353, 1989.

Zeevat,H., Klein, E., and Calder J., An Introduction to Unification Categorial Grammar, in Haddock N., Klein E. and Morill G. (eds) Edinburgh Working Papers in Cognitive Science, V.I: Categorial Grammar, Unification Grammar and Parsing, 1987.

Zue, V.et al, The Voyager Speech Understanding System: Preliminary Development and Evaluation, Proceeding ICASSP 90, pp73-76, 1990.