A Spoken Dialogue System for German Intercity Train Timetable Inquiries*

W. Eckert T. Kuhn H. Niemann S. Rieck A. Scheuer E.G. Schukat-Talamazzini
Universität Erlangen-Nürnberg,
Lehrstuhl für Mustererkennung (Informatik 5)
Martensstraße 3, 91058 Erlangen, Germany
wieland.eckert@informatik.uni-erlangen.de

Abstract

This paper focuses on the evaluation of the German SUNDIAL Demonstrator maintaining interactive conversations via microphone and telephone with users. The word recognizer was implemented by the University of Erlangen and currently obtains on our test set a word accuracy of over 92% in a speaker–independent task with perplexity 111.

We also participated in the design and implementation of the multilingual Dialogue Manager which is responsible for cooperative system behaviour. Over 100 prototype dialogues led to the current version which is continually being extended.

The overall system performance is tested with semi–naive users. Each subject faces four intercity train timetable scenarios, two of them are given, the others depend on the subjects' personal choice. Global system evaluation is done according to performance measures like contextual appropriateness, transaction success and dialogue completion rate.

1 Introduction

Many systems providing human computer interaction have been described, using pointing devices, speech input and/or output and even multimodal systems combining several input and output methods. Our prototype carries continuously spoken human machine dialogues utilizing speech input and output techniques, completely integrated and operational on a single workstation. The main components are: the acoustic phonetic recognizer front end (FE), the linguistic processor (LP) and the Dialogue Manager (DMan) — all of them part of the SUNDIAL research project.

Prototype systems have been developed by other partners for four languages with limited task domains such as flight schedules and train timetable inquiries. Each system carries out three principal functions: the interpretation of user utterances, the generation of system utterances and management of a coherent and natural dialogue. While the training of the FE and the interpretation in LP have to be language dependent, DMan was developed to operate on semantic units that represent language and domain independently.

*This work was partly funded by the Commission of the European Community DG XIII under ESPRIT contract P 2218 (SUNDIAL). Only the authors are responsible for the article.

After development of a fully operational prototype for all German partners, the evaluation of the system was started at Erlangen University. We decided to work with semi–naive subjects, who are familiar with computers but not knowing details about the dialogue system. The most prominent goal of this evaluation is the overall system performance, but detailed module evaluation is performed as well.

In this paper we present the first results of an early version of the integrated demonstrator setup at Erlangen University. Considering the weak points of the system (processing only the best string, preliminary version of the LP, slow processing speed) we found encouraging results.

2 Architecture

The hierarchical structure of our SUNDIAL system is shown in figure 1. An inhouse telephone is connected to the AD/DA–converter and the speech signal is digitized with 14 bit resolution at a sampling rate of 16 kHz. The AD/DA–converter is the only specialized hardware in the whole system.

Recognizer Mel cepstral features and their derivatives are processed in a SCHMM operating with stochastic bigram models. The principal phonetic subword unit of the HMM recognizer is the *polyphone* [11], representing a generalized context–dependent subword unit surrounded by arbitrarily large context. In contrast to triphones, the context is not artificially restricted to one symbol to the left and to the right; the context items may also include suprasegmental markers or even word boundaries. This ensures that large scaled contextual effects are properly statistically modeled. Design of the models and training of the HMM parameters is performed by the ISADORA system [10]. Currently, 1081 different words are modeled using 2991 subword units (SWU); 8674 probability density functions (PDF) are estimated resulting in a total of about 2 million parameters for the HMMs. Additionally 14 stochastic bigram models with perplexity ranging from 12 to 114 are predicted dialogue step dependently [1] and provide semantic restrictions for the recognizing process.

Parser The Linguistic Processor (LP) is an extended Tomita parser operating left–to–right and bottom–up [9] on the best string returned by the FE. Disjunctive constraints are represented in an efficient way to process the input. Using constraints replaces the copying and sharing of feature structures [2]. Finally the parse tree constructed

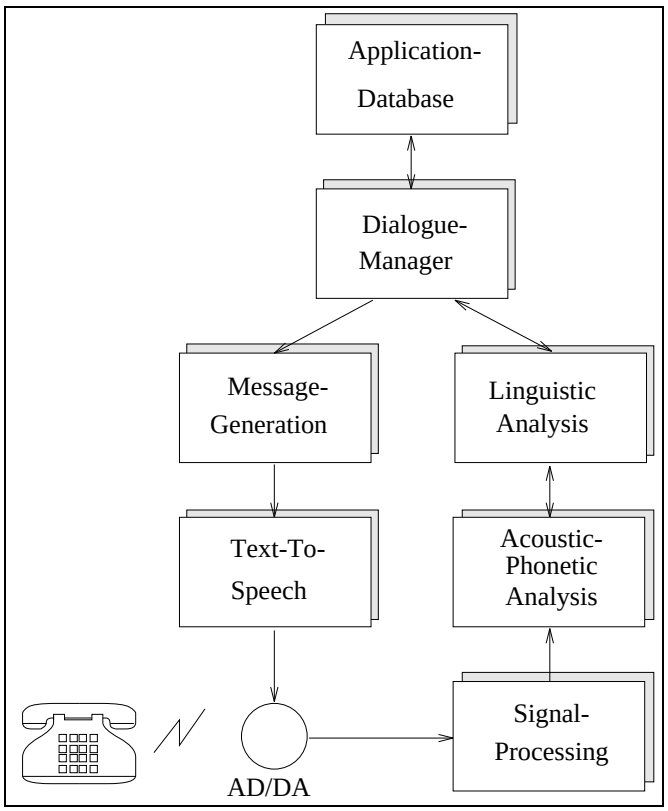

Figure 1: System architecture

from the word sequence is transformed into the semantic representation methodology developed in the project. We specify the semantic content of the utterance as well as the syntactic and morphologic information for further processing by the Dialogue Manager.

The semantic interface language (SIL) [7] provides a simple but flexible representation of utterances in terms of structured objects. Referencing objects, as well as the embedding of objects within objects, is facilitated by means of indexing: structure–sharing is achieved by co–referential indices. Objects (i.e. their labels and values) are intended to be theory–neutral, thus ensuring independence of linguistic representation and domain dependent processing. Any description of an utterance may consist of several *Utterance Field Objects* (UFOs), i.e. a structured collection of independent units representing largest coherent parsable sequences. Static and contextual interpretation of the UFO structures *dialogue*, *syntax* and *semantics* is performed by the DMan, which considers the linguistic history and the current dialogue status.

Dialogue Management The Dialogue Manager is designed and implemented in cooperation with partners from all countries participating in the SUNDIAL project, resulting in a language independent, multilingual system [8, 3]. Several different dialogue strategies (like parameter confirmation strategies, using defaults, telephone mode etc.) are implemented to allow a wide variety of different system reactions which are tested domain independently.

Management of spoken dialogues requires the interpretation of the user utterances and the generation of system utterances coherent with user utterances. The underlying interaction model consists of semantic, task and dialogue models [5]. Semantic descriptions of user utterances are processed, dealing with ambiguity, context–dependence and the hypothetical nature of the linguistic representation. In the task control module the set of parameters needed for an successful database enquiry are determined and further goal–directed requests for new parameters are initiated. In the dialogue interpretation process the semantic interpretations and the dialogue model are matched, deciding the subsequent steps of cooperative user interaction in the structured dialogue model. Using this partitioned interactional model, dialogue management is independent of the task and language of the service domain.

We have added to the system a special mode for requesting city names: the spelling mode. If in the same dialogue a city name is several times not confirmed (i.e. the user replies with a city name different from the understood name), the flexible DMan strategies request an explicit confirmation by the user utilizing the spelling mode. A special language model for the FE is selected, permitting only spelled city names. For that task the recognition module operated extremely well with a WA of nearly 100%: out of 194 city names only 3 were wrongly recognized. Two of them were spelt by a speaker with an Italian accent. Entering and leaving the spelling mode is transparent for the LP; neither the lexicon nor the grammar of the parser had to be extended.

Our installation is able to conduct information dialogues using high quality microphone input, as well as operating on a local telephone line, using the same parameterized components of the system.

3 Evaluation Methodology

Evaluation metrics are partitioned in two measure fields: isolated module and global system performance sections, corresponding with the glass box and black box evaluation methods.

Module Evaluation Evaluation of FE is performed using the well known measures *word accuracy* (WA), *word correctness* (WC) and *sentence recognition* (SR). They were calculated globally (on all dialogues) and on specific subsets of the trials. This is described below.

Usually the performance of the LP is measured using a set of test utterances spanning the expected grammatical and lexical coverage of the parser. Since we received a very early version of the LP from our partner, it was not evaluated according to that methodology. However, we counted the total number of parser failures, not distinguishing between parsing problems given a correct sentence and problems introduced due to misrecognition. Careful evaluation of the LP is not yet done.

The Dialogue Manager is designed using a set of over 100 different types of test dialogues, therefore ensuring the correct behaviour with all these dialogue prototypes. The *testfiles* define a subset of typical dialogue situations and incorporate the knowledge of conversational rules, contextual interpretation of utterances, dialogue strategies and database results.

System evaluation Apart from local module measures the system is tested as a whole with semi–naive users. The metrics *contextual appropriateness, turn correction ratio, transaction success* and *dialogue completion rate* are applied at the highest level of a dialogue system. They are described in table 1. Based on the conversational maxims [4], the contextual appropriateness is measured at the level of questions and answers. Turn corrections are divided into user corrections and system corrections, caused by misrecognition of the system and the user respectively. Turns which introduce problems and those which correct

them interrupt the flow of the dialogue without contributing new propositional content to it. They may, of course, make substitutions in the propositional content.

Measure	Definition
CA	*contextual appropriateness:* measures the appropriateness of a system utterance in its immediate dialogue context.
TCR	*turn correction ratio:* the ratio of all turns in a dialogue to those turns concerned with correcting troubles caused by misrecognition.
TS	*transaction success:* measures the system's success in generating the information the user requires, including pointing out that an answer does not exist and reacting cooperatively to inquiries with no concrete answer.
DC	*dialogue completion rate:* the fraction of all dialogues interactively finished without total system failure.

Table 1: System performance measures

Tagging of the user and system turns is performed by human experts who classify the appropriateness of utterances and the success of dialogues and also identify correction turns. Tags are counted automatically and summarized in a report along with the above mentioned FE performance measures.

Finally, the average length of a dialogue is calculated, showing the degree of interaction between system and user. Bad recognition rates tend to prolong the dialogues, but different user reactions to system utterances impose difficulties on the comparison of the average number of turns.

4 Experiments

20 semi-naive subjects (experienced in using computers, but not knowing details of the system) volunteered to perform dialogues with our system via microphone input. They were instructed by a person who supervised the tests but did not know any implementation details about the system and therefore could not predict the system behaviour. Each subject faced 4 intercity train timetable scenarios. Two of them were predefined, the other scenarios depended on the subjects' personal choice. Before starting their inquiries they had to define them by stating the places of departure and arrival and the departure time in a protocol. Each inquiry was tried until the system provided the required connection or spotted that no answer exists. After having carried out the dialogues the users had to fill in questionnaires describing their attitudes towards the system.

Apart from instructing the subjects, the supervisor kept a hand–written protocol during the sessions, giving us additional information about user and system reactions that were not automatically recorded by the system in a protocol file. The hand–written protocol, for instance, tells us what the subject actually said whereas the protocol file only records what was recognized by the acoustic front end processor. After finishing the experiments all information contained in the hand-written protocol was inserted into the protocol file and specifically marked in order to facilitate the extraction of evaluation material.

5 Results

System evaluation The test material was first analyzed to find the total failures of the system. Then the dialogues that do not lead to a system shutdown are evaluated according to the already described metrics. The evaluation of the questionnaires showing the users' judgments is not presented in this paper.

The total number of 255 started dialogues consists of three roughly equal parts (table 2). One third of the dialogues were finished with the correct solution. The second third of the dialogues terminated indicating that the dialogue progress is not good enough to expect successful completion. Finally another third were terminated unexpectedly due to a severe bug, causing an immediate system shutdown.

Measure: DC	no.	%
failed dialogues	94	36.8%
successful dialogues	79	31.0%
spotting "difficulties"	82	32.2%
total successful	161	63.1%

Table 2: Dialogue completion rate

The evaluation focused on 79 dialogues that were successfully finished and was carried out by means of the accuracy measurements mentioned in table 1.

Nearly all system utterances were judged to be appropriate (table 3), i.e. bringing the dialogue forward, even if not all parameters in the users utterance are recognized correctly. If no progress is observable within a number of turns, the system itself terminated the dialogue announcing the shutdown due to a high rate of misunderstanding.

Measure: CA	no.	%
AP appropriate	635	98%
IA inappropriate	11	2%
AI not agreed AP/IA	0	0%
TF total failure (ruled out)		
IC incomprehensible	0	0%

Table 3: Contextual appropriateness of system utterances

The transaction success rate (TS) is subdivided into the parts *successful, successful with constraint relaxation, successful spotting that no connection exists* and *failure*. The German task of train timetable inquiries implies a large number of solutions with relaxed parameters, e.g. proposing a train at 17:47 when one at 18:00 was requested. The TS results are shown in table 4.

Measure: TS	no.	%
S successful	5	6%
SC relaxed constraints	63	80%
SN announcing no solution	11	14%
F failure	0	0%

Table 4: Transaction success of the dialogues

The dialogue completion rate is calculated by counting the finished dialogues and the started dialogues. Even the dialogues announcing the system shutdown without delivering a solution are counted as *finished* because the system still produces an answer. Details are shown in table 2.

Finally, the average length of a dialogue was calculated, describing the capabilities of the system to conduct a longer interaction. Table 5 indicates that the premature shutdown arose in the beginning or the middle of a dialogue. Correctly finished dialogues averaged at about eight user turns, caused by mild misrecognitions. Some subjects judged the dialogues as too lengthy.

Dialogue length	user turns
failed dialogues	3.6
successful dialogues	8.0
spotting "difficulties"	6.1

Table 5: Average number of user turns per dialogue

Module evaluation Apart from the overall evaluation of the whole system, we concentrated on inspection of the FE and DMan performance. Additionally we examined the parser success rate.

Table 6 summarizes the FE results and shows performance with an overall word accuracy of 66.5%. In those dialogues which were terminated by the dialogue manager announcing too serious difficulties, the WA dropped by more than 10 points, justifying the DMan's decision. This, too, is reflected by the sentence recognition rate (SR), falling below 30%.

FE performance	WA %	WC %	SR %
all dialogues	66.5%	77.2%	47.1%
successful dialogues	70.4%	81.7%	57.6%
spotting "difficulties"	54.6%	68.9%	29.7%

Table 6: Acoustic Front End performance measures

The recognizer performed significantly different in out test set and the user trials. This gap is caused mainly by the characteristics of spontaneous speech vs. read speech and local dialects of the subjects.

When evaluating the LP performance using a best string input, one has to distinguish between parser failures due to receiving incorrect utterances from the FE and correctly recognized utterances which are out of coverage of the LP. However, in this first evaluation phase we summarize both effects under the title *parser failed* in table 7.

LP performance	total utterances	parse failed	%
all dialogues	1476	699	47.4%
successful dialogues	635	243	38.3%
spotting "difficulties"	500	366	73.3%

Table 7: Linguistic Processor performance measures

Again, the significant increase of the error rate justify premature dialogue closings.

6 Conclusion

We described the evaluation of a prototype of our dialogue system, consisting of the main modules *acoustic front end, linguistic processor* and *dialogue manager*.

Apart from presenting module–based results, we focused on the evaluation of the completely integrated system. Due to a major bug in the software some dialogues were not finished but all the remaining dialogues showed the characteristics of a *dialogue* system, i.e. asking questions and reacting to the users utterances. Half of them were closed politely by the dialogue manager announcing the premature end due to severe understanding problems. The other half finished satisfactorily, giving the information the user wanted.

Recently, we isolated the above mentioned bug and received a newer release of the linguistic processor. The grammar was extended to handle in a more robust way the syntactic (and semantic) incorrect word sequences that resulted from misrecognition of single words; this new version of the grammar yields a partial description of the input in this case. Further tests using the new version showed promising results. A second evaluation phase with all updates incorporated into the system is planned for the near future.

7 Acknowledgments

We acknowledge the cooperation with the other SUNDIAL partners: CAP Gemini Innovation, CNET, CSELT, Daimler–Benz (Ulm), Infovox, IRISA, LOGICA, Politechnico di Torino, Sarin, Siemens (Munich) and University of Surrey.

We wish to thank our German partners: G. Niedermair (Siemens, Munich) provided the linguistic processor, the colleagues with Daimler–Benz (Ulm) performed the telephone adaptation and delivered the speech synthesizer.

References

[1] F. Andry. Static and dynamic predictions: a method to improve speech understanding in cooperative dialogues. In *Proc. Int. Conf. on Spoken Language Processing*, pages 635–638, Banff, 1992.

[2] H.U. Block and L.A. Schmid. Using disjunctive constraints in a bottom-up parser. In *Konvens 92*, pages 169–177, Berlin, October 1992. Springer.

[3] W. Eckert and S. McGlashan. Managing Spoken Dialogues for Information Services. In *Proc. European Conf. on Speech Technology*, Berlin, Germany, 1993 (this issue).

[4] H.P. Grice. Logic and Conversation. In *Syntax and Semantics 3: Pragmatics*, New York, 1975.

[5] B. Grosz and C. Sidner. Attention, Intensions, and the Structure of Discourse. *Computational Linguistics*, 12(3):175–204, July–September 1986.

[6] T. Kuhn, H. Niemann, E.G. Schukat-Talamazzini, W. Eckert, and S. Rieck. Context-Dependent Modeling in a Two-Stage HMM Word Recognizer for Continuous Speech. In J. Vandewalle, R. Boite, M. Moonen, and A. Oosterlinck, editors, *Signal Processing VI: Theories and Applications*, volume 1, pages 439–442. Elsevier Science Publishers, Amsterdam, 1992.

[7] S. McGlashan. A Proposal for SIL. Technical report, Sundial WP6 (unpublished), 1991.

[8] S. McGlashan, E. Bilange, N. Fraser, N. Gilbert, P. Heisterkamp, and N. Youd. Managing Oral Dialogues. In *6th International Workshop on Natural Language Generation*, Trento, Italy, April 1992.

[9] G. Niedermair. Linguistic Modelling in the Context of Oral Dialogue. In *Proc. Int. Conf. on Spoken Language Processing*, pages 635–638, Banff, 1992.

[10] E.G. Schukat-Talamazzini and H. Niemann. ISADORA — A Speech Modelling Network Based on Hidden Markov Models. *Computer Speech & Language*, 1993 (to appear).

[11] E.G. Schukat-Talamazzini, H. Niemann, W. Eckert, T. Kuhn, and S. Rieck. Acoustic Modelling of Subword Units in the ISADORA Speech Recognizer. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, volume 1, pages 577–580, San Francisco, 1992.