In H. Niemann, *Progress and Prospects of Speech Research and Technology*, infix, 1994.

Prosodic Dialog Control in EVAR

H. Niemann, W. Eckert, A. Kießling, R. Kompe, T. Kuhn, E. Nöth,
M. Mast, S. Rieck, and E.G. Schukat–Talamazzini,
Universität Erlangen-Nürnberg,
Lehrstuhl für Mustererkennung (Informatik 5),
Martensstr. 3, 91058 Erlangen, F.R. of Germany
e-mail: kompe@informatik.uni-erlangen.de

A. Batliner
L.M.-Universität München, Institut für Deutsche Philologie,
Schellingstr. 3, 80799 München, F.R. of Germany

Abstract

The domain of the speech recognition and dialog system EVAR is train time table inquiry. We observed that in real human–human dialogs when the officer transmits the information, the customer very often interrupts. Many of these interruptions are just repetitions of the time of day given by the officer. The functional role of these interruptions is determined by prosodic cues only. An important result of experiments with naive persons is that it is hard to follow the EVAR system giving the train connection via speech synthesis. In this case it is even more important than in human-human dialogs that the user has the opportunity to interact during the answer phase. Therefore we extended the dialog module to allow the user to repeat the time of day and we added a prosody module guiding the continuation of the dialog by analyzing the intonation contour of the utterance.

1 Introduction

Dialog systems for information retrieval are potential applications for human–machine communication. In human–human dialogs, it is often the case that parts of the information just given by the speaker are repeated by the partner. For example, in train time table inquiries it can be observed frequently that the customer repeats the times of arrival or departure just given by the officer. Frequently only the intonation of this repetition of the time–of–day shows the intention of the customer and thus governs the continuation of the dialog.

In the scenario of our speech understanding and dialog system EVAR (an experimental automatic information system for train time table inquiries) the transmission of these times is a pivot point. The most convenient way to generate an answer in this application is a printed time table. However, in the case of information retrieval via telephone, the answer has to be generated by a speech synthesis system. In many applications such as in ours the answer can be quite lengthy, especially when there is a transfer. Even if one is accustomed to the unnatural synthetic voice, it is often hard to follow the answer given in one piece. A possible, but certainly not user friendly solution, would be to generate the answer slowly and with many pauses. A

better approach is to allow for an interruption whenever the user didn't understand part of the information.

Of course, in the case that the user is allowed to interrupt the answer given by the system, a user–friendly system should be able to react adequately (cf. [Wai88]). Let us consider the following dialog: *officer*: *"... leaves Ulm at 17 23."* *customer*: *"17 23./?"*. In the case of a rising intonation (denoting a question: '?') the officer — or the system, respectively — has to repeat the time–of–day, because the customer wants to hear the time again. In the case of a falling intonation (denoting a confirmation: '.') no specific reaction is necessary and the system can give the next part of the information.

Following the ideas of Nöth (presented in [Nöt91]), this paper describes how the dialog module of EVAR has been extended to allow for such repetitions of the time–of–day by the user and how adequate reactions by the system based on the hypotheses computed by a prosody module are implemented. The paper is organized as follows: First (section 2) we give an overview of the speech recognition and understanding system EVAR. In section 3 the dialog module of EVAR without prosody is described, including results of recent experiments with naive subjects using EVAR. Motivated by these and by the observation of real human–human dialogs (section 4) we extended the dialog module and added a prosody module to the system, which is described in the final part of the paper (section 5). The paper concludes with a discussion.

2 The Speech Understanding System EVAR

The speech understanding and dialog system EVAR (the acronym stands for the German words for "to recognize" — "to understand" — "to answer" — "to ask back") is an experimental automatic travel information system in the domain of German *InterCity* train time table inquiries. This section will give an overview; for a more detailed description of the EVAR system see Mast et al. [MKE+94]. Figure 2 shows the structure of EVAR. The two main components are the linguistic analysis and the acoustic processing.

Input to the system is continuous German speech, which is recorded and digitized with 16 kHz and 14 bits with a DeskLab from Gradient directly connected to the work station where the system is running. No other special hardware is used. In the current version, output of the speech recognition component is the best matching word sequence, but word hypotheses graphs can be used as well. The generation of word sequences is based on Hidden Markov Models (see [STNE+93]).

For the representation of knowledge the semantic network system ERNEST is used (cf. [NSSK90]). All knowledge needed for the speech understanding process and for the dialog is embedded within a single semantic network using the same representation language. Thus it is easy to propagate restrictions from all levels to support the recognition process. Nevertheless the knowledge base is easy to extend and to modify because of its modularization into *levels of abstraction*. These levels are connected with *concrete (con)* links. Within a level of abstraction concepts are connected using *part–of* and *specialization (spec)* links. In the semantic network the following modules are integrated (see Figure 2):

- The *word hypotheses* module builds up the interface between speech recognition and linguistic analysis. Word hypotheses restricted to the linguistic and task–specific expectations (which depend on the actual state of the analysis) are requested from the acoustic–phonetic front–end.

- In the *syntactic* module, syntactic constituents and special dialog constructions are represented but not the order of the constituents on the sentence level. In German it is

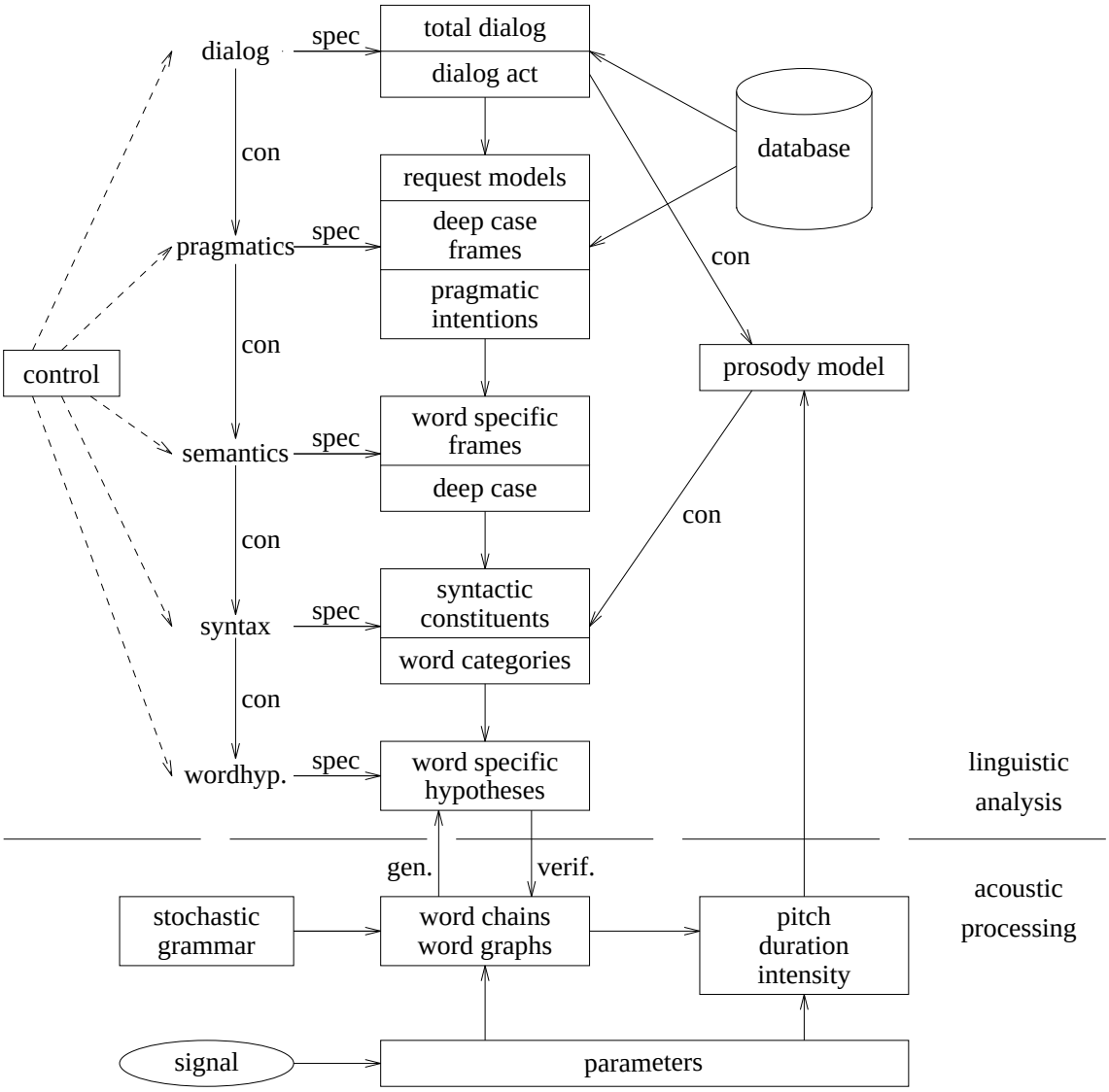

Figure 1: The speech recognition and dialog system EVAR.

characteristic for spontaneous speech that the order can be rather free.

- In the *semantic* module, verb and noun frames with their deep cases according to Fillmore's deep case theory are modelled (cf. [Fil68]).
- The *pragmatic* module represents task–specific knowledge.
- The *dialog* module models possible sequences of dialog acts (cf. section 3). Further, natural language system answers are generated and handed to the text–to–speech system SPRAUS from Daimler Benz.
- The *prosody* module models the suprasegmentals of the user utterances (cf. section 5.3).

In addition to the provision of the knowledge representation scheme, ERNEST provides mechanisms to use this knowledge for the analysis process. The problem–independent procedural semantics of the network language allows a flexible control of the analysis process, which is alternating between bottom–up and top–down search. The A^*-Algorithm in combination with problem–dependent judgement vectors is the basis of this control. The ultimate goal of the analysis is the instantiation of a sequence of dialog–level concepts until all the parameters for a database request are known. A concept can be instantiated when one out of a collection

of predefined subsets of concepts could be instantiated to which it is connected via *concrete* or *part–of* links.

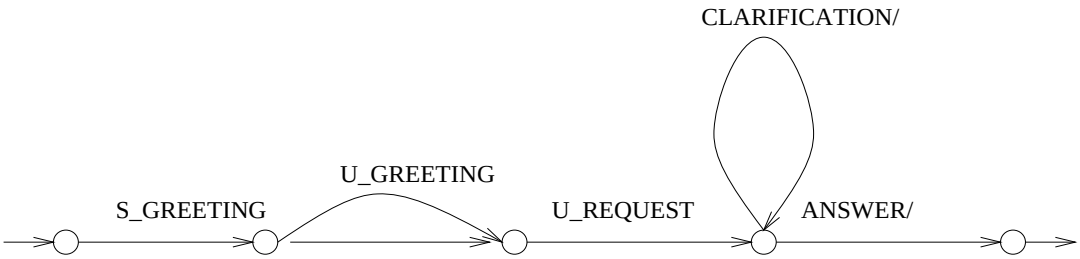

Figure 2: Recursive transition network representing the dialog model implemented in EVAR.

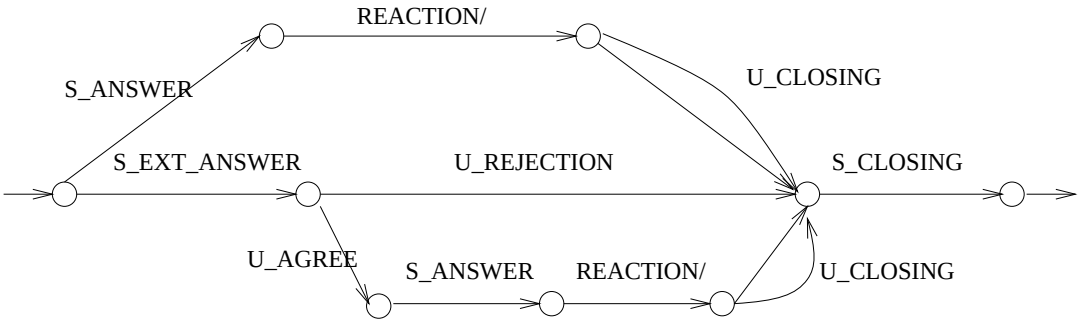

Figure 3: The ANSWER/ subnet.

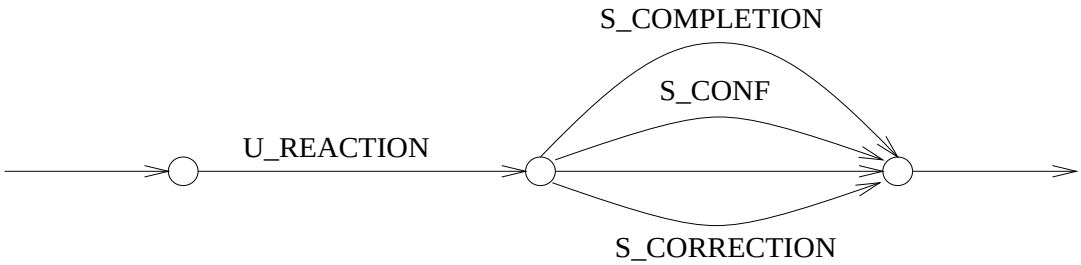

Figure 4: The REACTION/ subnet (cf. section 5).

For experiments with naive users as described in section 3.2, it is important to give correct train connections. Otherwise the users often don't take the experiment seriously as has been pointed out in [HK89]. Therefore the EVAR system is connected to the on–line version (HaFas, developed by HaCon GmbH) of the official German train time table. The time needed for the retrieval of one inquiry is far less than one second.

3 The Dialog Module without Prosody

A user utterance has to be interpreted syntactically, semantically and pragmatically as well as in the dialog context. The latter comprises both the knowledge about what kind of utterances may follow each other, and the consideration of the dialog history in order to be able to resolve anaphoric references and to focus the analysis on the expected answer. In the following, an

overview of the dialog module is given and experiments with the system are presented (for more details see [MKK+92, Mas93]).

3.1 The Dialog Model

In a user–friendly system the user should have the possibility of talking to the system without extensive restrictions, i.e., almost as if he were talking to an information officer. The dialog model must therefore represent all expected sequences of dialog acts which are typical in this special situation. From a corpus of real human–human dialogs (cf. [HUKW86]), a model was extracted containing all the sequences of dialog acts observed in the corpus which are relevant for human–machine communication. Figure 2 shows a recursive transition network representing the dialog model implemented in EVAR. One edge corresponds to one dialog act or refers to a subnet (indicated by a slash). The prefixes S_, U_ indicate that the dialog act corresponds to a **s**ystem or a **u**ser utterance, respectively. The subnet for clarification will not be discussed in this paper. Figure 3 shows the subnet for the answer phase (ANSWER/). The subnet "REACTION/" (Figure 4) contains the extensions to the dialog model relevant for this paper. It is described below (section 5) and was not implemented in the version of the system that was used for the experiments described in this section.

Each dialog act is modeled by a set of pragmatic, semantic and syntactic concepts representing what the user is expected to utter. The properties of the concepts and the current dialog state are used to identify the actual dialog act.

After the greeting, the user requests information. If the information that is necessary for giving an answer is not contained in the user's request, the system starts a clarification dialog which is not the topic of this article.

The user utterances have to be syntactically and semantically complete or they have to be incomplete in such a way that they can be completed by taking parts of prior utterances. For the answer generation, sentence masks are used for each dialog act. The following examples for the different dialog phases are translated into English (the abbreviations of Figures 2 and 3 are given in parentheses):

S: (S_GREETING) Hello. This is the Automatic Travel Information System EVAR.

U: (U_REQUEST) Good morning. I want to go to Hamburg tomorrow in the afternoon.

S: (S_EXT_ANSWER) You can take the train at 14h15. You switch trains in Würzburg at 15h20. You will arrive in Hamburg at 19h10. Do you want a later train?

U: (U_REJECTION) No thanks.

S: (S_CLOSING) Thank you for calling the Automatic Travel Information System, good bye.

3.2 Experiments with naive subjects

With this system, experiments with 15 naive subjects were conducted (cf. [Mas93]): The users were asked to take the part of the customer in at least four different scenarios. Two scenarios were given and were the same for every user; the other scenarios were created by the users themselves. The experiments were conducted in a quiet office environment using headphones. A total of 82 dialogs were recorded.

The word recognizer was trained on 7900 domain–specific read sentences from 79 speakers (100 sentences each). The lexicon contained 1081 words; a bigram language model of perplexity 111 was used. The word accuracy in the above mentioned experiments was 73.7% (79.9% of

the words and 38.2% of the sentences were correct). For comparison: on read sentences (not contained in the training set) a word accuracy of 92% was achieved. This difference is due to two facts:

1. The utterances contained typical spontaneous speech phenomena such as false starts, repetitions, filled pauses and non–speech events (cf. [O'S92, SL92, SBD92]) which are not yet modeled by the word recognizer (compare [BMSP92]).

2. Even if the words uttered are modeled by the word recognizer, a number of errors may occur since the recognizer was trained on read speech and there are many differences between read and spontaneous speech (compare [DZ90, DZ92, BKK+94]).

Forty of the 82 dialogs were completed successfully, i.e., the system provided the correct train connection. Eight dialogs were completed but the system didn't provide the information the user asked for due to an incorrect analysis of parameters needed for the database request. The rest of the dialogs were not completed due to memory limitations, repeated misunderstandings of utterances or the user giving up the dialog.

To assess user satisfaction after each session the users were asked to answer a questionnaire. One question was: "Would you use such a system if it would be available via telephone?" Twelve of the 15 users answered that they would use it with a few improvements, e.g., that the system would be faster and the answers would be presented slower, with details and a possibility for repetition. Five subjects answered with "No" (multiple answers were possible), e.g., because it was too slow or because they found such a machine not very social. Another question was if it took much effort to learn to communicate with the system. Two people answered that it took some effort to learn it and 13 people answered that it took no effort.

For the experiments, the EVAR system was run on a DEC station 5000/200. The time for the generation of the word hypotheses was 4.2 times real time. The average CPU time for the linguistic analysis and interpretation in the dialog context for one utterance was 44 seconds. The average time to complete a dialog was 9.5 minutes.

4 Dialog Guiding Prosodic Signals

Since the goal of EVAR is to conduct dialogs over the telephone, the system answer is generated by a speech synthesis system. As has been motivated in sections 1 and 3.2, the system should allow for user interruptions and react adequately to them. In order to derive a formal scheme for this, we investigated a corpus of 107 "real–life" train time table inquiry dialogs recorded at different places, most of them conducted over the phone. Ninety–two dialogs concerned train schedules; the rest had other topics such as fares.

The most important question in the context of this paper is how often and in which way during the answer phase the prosody of a user interruption alone controls the following action of the officer. In this section we will summarize the main results of this investigation. For further details see [BKK+92].

In the following, only the 92 dialogs concerning train schedules are considered. Among these there are 215 utterances in which the customer repeats the time of arrival or departure given by the officer (a total of 227 repetitions of the time–of–day), i.e., more than two repetitions per dialog on the average. In all but 3 cases, the repetition concerned the time–of–day the officer had just given before. There are two forms of time–of–day expressions possible in German: with or without the word *Uhr* which means *o'clock* (e.g., *"17 Uhr 23"* or *"17 23"*).

By repeating the time–of–day, the customer expresses different aims, i.e., he wants to give the officer different kinds of information. The reaction of the officer and thus the continuation

System answer: *"... In München sind Sie dann um 17 Uhr 32."* *"... You'll arrive in Munich at 5 32 p.m."*					
RTD		intonation		system reaction	
no utterance		——		——	
wrong repetition		——		correction	*('Nein, um 17 Uhr 32.')*
complete & correct		rising	*('17 Uhr 32?')*	confirmation	*('Ja, um 17 Uhr 32.')*
		continuation rise	*('17 Uhr 32–')*	——	
		falling	*('17 Uhr 32.')*		
correct & incomplete	only minutes	rising	*('32?')*	confirmation	*('Ja, um 17 Uhr 32.')*
		continuation rise	*('32–')*	——	
		falling	*('32.')*		
	only hours	rising	*('17 Uhr?')*	completion	*('17 Uhr 32.')*
		continuation rise	*('17 Uhr–')*		
		falling	*('17 Uhr.')*	——	

Table 1: The reaction scheme for repetitions of the time of day (RTD) within the dialog system EVAR. (The word "*Uhr*" means "*o'clock*", "*nein*" = "*no*", "*ja*" = "*yes*", "*um*" = "*at*".)

of the dialog is governed by the specific kind of information which is mostly expressed by the intonation. We observed three different functional roles of the repetition of time–of–day: *confirmation*, *question* and *feedback*:

- Using a **confirmation**, the customer wants to signal the officer that he received the last information, e.g., the time of arrival. Usually, the intonation ($F0$–contour) at the end of such an utterance is falling.
- With a **question** the customer signals the officer that he did not understand, i.e., that he did not get the time–of–day completely or that he just wants to ask the officer to confirm the correctness. The prototypical $F0$–contour at the end of such an utterance is rising.
- By using a **feedback**, the customer usually wants to signal the officer *"I'm still listening"*, *"I got the information"*. It is usually characterized by a level or slightly rising $F0$–contour (continuation rise).

In our material, in 100 of the 227 repetitions of the customer the reaction of the officer (confirmation of the correctness, correction or completion of the time–of–day) was governed by nothing but the intonation of the customer. In the remaining cases, there were other indicators such as *Wh*–words (e.g., "*When at five seventeen?*"). In 64 of the 100 cases, the time–of–day occurred isolated; the other cases contained words that didn't indicate the desired response such as "*Leave Munich at five seventeen*".

From the corpus we developed a scheme (see Table 1) showing the reactions of the officer depending on the intonation of the repetition of time–of–day by the customer. The dialog module of EVAR, which in our application plays the role of the officer, was extended on the basis of this scheme (cf. section 5).

In the scheme it was not only taken into account whether the customer repeated the time–of–day correctly and completely (note, that also the expression "*21 Uhr*" is complete, if the officer said this before), but also if she/he repeated the time–of–day incompletely (but correctly) or incorrectly (see Table 1, column 1). Column 2 of Table 1 shows the type of the intonation contour of the customer utterance, which was classified manually by an expert listening to the signals. The entries in the first two columns completely determine the reaction of the officer, which can be correction, completion, confirmation or no special reaction, i.e., the officer proceeds as if the user had said nothing. Looking at the rows of Table 1 the first one ("no

utterance") seems to be trivial: if the user does not utter anything, then there is no officer reaction. However, this case also has to be explicitly taken into account in our system (cf. section 5). If the repetition is incorrect, the intonation contour is irrelevant and the officer corrects the customer in any case. If the time–of–day is repeated correctly and completely or if the minutes alone are repeated correctly an interrogative contour of the customer utterance provokes an officer reaction, which is confirmation; fall or continuation rise both indicate that the customer believes that he understood the officer utterance. When the customer repeats the hour alone (and the officer has uttered a time–of–day containing hour and minutes), then in the case of rise or continuation rise, we observed that the officer completes the customer utterance by either repeating the minutes alone or by repeating the complete time–of–day; in the case of a falling contour the customer confirms the officer utterance so that the officer shows no special reaction. (In our data there was only one case of correct and incomplete repetition of the hour alone with a continuation rise; we believe that both completion or no special reaction by the officer would be plausible.)

5 The Dialog Module with Prosody

To cope at least partly with the problems mentioned in section 3.2, we extended the dialog module of EVAR and added a prosody module to the semantic network such that the repetitions of the time–of–day as described in section 4 are modeled.

5.1 Classification of Sentence Modality

In order to be able to model the potential user reactions, we have trained a Gaussian classifier to distinguish the three classes of sentence modality (i.e., fall, rise and continuation rise), that are mapped onto the functional roles of the repetition of time–of–day (i.e., confirmation, question, and feedback). From the automatically computed $F0$–contour (cf. [KKN+92]) the following four features were extracted considering only the voiced frames (non–zero values): the slope of the regression line of the whole and of the last two voiced regions of the $F0$–contour, and the differences between the offset (the $F0$-value of the last voiced frame) and the values of each of the two regression lines at this offset position. On a database of 322 isolated time–of–day utterances from four speakers a recognition rate of 87.5% was achieved in leave–one–out experiments.

5.2 Extension of the Dialog Module

The repetitions of the time–of–day of the user and the appropriate system reactions have been represented in the dialog module by introducing a new subnet (REACTION/, see Figure 4). After the system has given the answer (i.e., a train connection) the user has the opportunity to repeat the time–of–day previously uttered by the system (edge U_REACTION — user reaction — in Figure 4). In the current implementation there is always a signal recorded for a fixed amount of time. Therefore silence is interpreted as a user reaction as well (see Table 1). After the user reaction the system has four alternatives: completion (S_COMPLETION), correction (S_CORRECTION), confirmation (S_CONF) or no special reaction (empty edge). After each of these alternatives it proceeds with the closing (S_CLOSING, Figure 3). Which one of these alternatives is chosen depends on the reaction scheme of Table 1, which is implemented in the control module of EVAR.

Each of these dialog steps is implemented as a concept in the semantic network of EVAR. The concept for the user reaction is connected via *concrete* links to the following concepts (see

also section 5.4):

- a concept representing silence. During analysis at first it is tried if this concept can be instantiated, by applying an attribute, which checks if there was only silence recorded.
- a concept, which is responsible for the syntactic and semantic analysis of time–of–day expressions. It is instantiated if the instantiation of the silence concept failed. This concept itself has *part–of* and *concrete* links to other concepts. With this the search space for the linguistic analysis and word recognition is restricted to time–of–day expressions.
- a concept of the prosody module representing sentence modality (cf. section 5.3).

5.3 The Prosody Module

In the current system the classification of the intonation contour is done with the Gaussian classifier described in section 5. Implemented is also an alternative approach comparing the actual intonation contour with a set of prototypical F0–contours via dynamic time warping. This might give better results, since the intonation contour depends very much on the corresponding word chain, especially on the number of syllables in the utterance and the position of the accent. However, constructing a set of prototypes is very time consuming and we cannot yet report any recognition results.

At present the prosody module integrated in the semantic network consists of one concept for sentence modality and a set of attributes defining knowledge about the intonation of time–of–day utterances, and another concept whose attributes perform the classification and establish an interface to the (so–far) external process computing the F0–contour. The prosody concepts are linked to the dialog module and to the syntax module. The links to the dialog module had to be established to allow for a prosodically guided dialog control. The links to the syntax module were necessary since in the case of classification, where the computed F0 contour and prototypical contours are matched via dynamic time warping, the prosody module has to have access to the word chain underlying the semantic interpretation, so that prototypes can be chosen depending on the number of syllables in the recognized word chain.

In order to use prosody to control the dialog a decision is necessary about the type of the intonation contour. Thus the utterance is classified by the classifier and one instance of the sentence modality concept is created corresponding to the most probable class of intonation contour (e.g., rise). Since we are working on the use of other prosodic information (cf. below) we designed the concepts in such a way that they can be used in a more flexible manner. For example, for the disambiguation of the attachment of prepositional phrases or the boundary between main and initive clause one would need hypotheses for prosodic phrase boundaries (i.e., several scored instances of concepts modeling prosodic phrase boundaries) and hypotheses for different intonation contours at each predicted boundary so that the control module can search for the "optimal" interpretation integrating all levels of knowledge.

5.4 The Analysis Process

In the previous section, we described the structure of the extended knowledge base. In the following we will sketch the analysis steps within the parts of EVAR corresponding to the extensions of the dialog model described in section 5 (subnet REACTION/, see Figure 4) As pointed out in section 5, in the dialog act U_REACTION a signal is recorded in any case.

Then a separate module determines if the signal only consists of silence (this corresponds to the first row of Table 1). In that case a "silence word hypothesis" is handed to the linguistic analysis and no further word recognition has to be done. Then, the silence concept (cf. section 5)

is instantiated during linguistic analysis. After this the dialog ends directly with the closing (S_CLOSING).

If there is not only silence in the signal, the word recognizer computes the best word chain. Since the word recognizer is integrated via procedure call, we can easily use dialog act–dependent language models. If the user interrupts, the vocabulary and the bigram language model are restricted to time–of–day expressions, which can be [hour], [hour] [minute], [hour] *Uhr* [minute], or just [minute]. The word accuracy on a database of in total 200 time–of–day expressions read by two female and two male "naive" speakers is about 82%. Despite the reduced vocabulary and perplexity compared to the results mentioned in section 3.2, the accuracy is lower because the similarity between the allowed words is much higher.

Now the best word chain is semantically interpreted as a time–of–day expression. As a result, the concept for the analysis of time–of–day expressions is instantiated. This expression is compared to the last time–of–day given by the system. Six cases can be distinguished:

1. the user did not utter a time–of–day expression but the language model forced the recognizer to recognize one
2. the user misunderstood the system and repeated the wrong time–of–day expression
3. the user utterance was misrecognized by the word recognizer
4. the utterances of the system and of the user agree semantically
5. the user only repeated the minute expression
6. the user only repeated the hour expression

In the first three cases, the intonation contour is not classified, i.e., the concept for sentence modality is not instantiated. The dialog proceeds with the dialog act S_CORRECTION, i.e., the system corrects the user and repeats the last time–of–day (see Table 1, row 2).

In the other three cases prosody is used for the selection of the next dialog act and the intonation contour is classified as described in section 5.3. Then the concept for sentence modality and user reaction are instantiated, and the dialog proceeds with the next dialog act (confirmation, correction or completion) according to the scheme in Table 1.

6 Discussion and Future Work

Already in [Lea80] and [Vai88] the integration of a prosody module into automatic speech understanding (ASU) systems is discussed. Lea even proposed a control module very much driven by prosody. To our knowledge, however, this paper presents the first dialog system partly guided by prosodic information. The system is still at an experimental stage, i.e., the user, so far, cannot really interrupt a system utterance, but after each system answer the user gets the chance to react. Up to now the train connection is given within a single utterance. We are working on splitting the system answer into small pieces, each uttered separately allowing for a "quasi–interruption" by the user. These restrictions do not affect the main goal of the work leading to this paper, i.e., the development of principal methods for integrating a prosody module into the overall system and getting it to interact with the other system components, especially to guide the progress of the dialog. However, due to these restrictions we were not yet able to conduct realistic experiments with the extended EVAR.

In the future, we plan to take into account different possibilities of accentuation as well as non–isolated repetitions of time–of–day. In addition, we have begun to work on the integration of prosody at other levels of our ASU system. The integration of accent information into a word

recognition module is under investigation, and the use of prosodic phrase boundaries during syntactic parsing and for re–scoring the n–best sentence hypotheses is being explored. First results concerning the recognition of prosodic phrase boundaries and accents are presented in [KBK+94] and [KKB+94] (see also [OWV93], [WH92] and [Hub89]).

7 Acknowledgements

This work was supported by the German Ministry for Research and Technology (*BMFT*) in the joint research project ASL and by the German Research Foundation (DFG). Only the authors are responsible for the contents of this paper.

References

[BKK+92] A. Batliner, A. Kießling, R. Kompe, E. Nöth, and B. Raithel. Wann geht der Sonderzug nach Pankow? (Uhrzeitangaben und ihre prosodische Markierung in der Mensch-Mensch– und in der Mensch-Maschine-Kommunikation). In *Fortschritte der Akustik — Proc. DAGA '92*, volume B, pages 541–544, Berlin, 1992.

[BKK+94] A. Batliner, R. Kompe, A. Kießling, E. Nöth, and H. Niemann. Can you tell apart spontaneous and read speech if you just look at prosody? In A. Rubio, editor, *New Advances and Trends in Speech Recognition and Coding*, NATO ASI Series F (to appear). Springer–Verlag, Berlin, Heidelberg, New York, 1994.

[BMSP92] J. Butzberger, H. Murveit, E. Shriberg, and P. Price. Modeling Spontaneous Speech Effects in Large Vocabulary Speech Recognition Applications. In *Speech and Natural Language Workshop*. Morgan Kaufmann, 1992.

[DZ90] N.A. Daly and V.W. Zue. Acoustic, Perceptual, and Linguistic Analyses of Intonation Contours in Human/Maschine Dialogues. In *Int. Conf. on Spoken Language Processing*, pages 497–500, Kobe, 1990.

[DZ92] N. Daly and V. Zue. Statistical and Linguistic Analyses of F0 in Read and Spontaneous Speech. In *Int. Conf. on Spoken Language Processing*, volume 1, pages 763–766, Banff, 1992.

[Fil68] Ch. Fillmore. A case for case. In E. Bach and R. T. Harms, editors, *Universals in Linguistic Theory*, pages 1–88. Holt, Rinehart and Winston, New York, 1968.

[HK89] L. Hitzenberger and H. Kritzenberger. Simulation Experiments and Prototyping of User Interfaces in a Multimedial Environment of an Information System. In *Proc. European Conf. on Speech Communication and Technology*, volume 2, pages 597–600, Paris, September 1989.

[Hub89] D. Huber. A statistical approach to the segmentation and broad classification of continuous speech into phrase-sized information units. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pages 600–603, Glasgow, 1989.

[HUKW86] L. Hitzenberger, R. Ulbrand, H. Kritzenberger, and P. Wenzel. FACID Fachsprachlicher Corpus informationsabfragender Dialoge. Technical report, FG Linguistische Informationswissenschaft Universität Regensburg, 1986.

[KBK+94] R. Kompe, A. Batliner, A. Kießling, U. Kilian, H. Niemann, E. Nöth, and P. Regel-Brietzmann. Automatic Classification of Prosodically Marked Phrase Boundaries in German. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 173–176, Adelaide, 1994.

[KKB+94] A. Kießling, R. Kompe, A. Batliner, H. Niemann, and E. Nöth. Automatic Labeling of Phrase Accents in German. In *Int. Conf. on Spoken Language Processing* (to appear), Yokohama, September 1994.

[KKN+92] A. Kießling, R. Kompe, H. Niemann, E. Nöth, and A. Batliner. DP-Based Determination of *F0* contours from speech signals. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages II–17–II–20, San Francisco, 1992.

[Lea80] W. Lea. Prosodic Aids to Speech Recognition. In W. Lea, editor, *Trends in Speech Recognition*, pages 166–205. Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1980.

[Mas93] M. Mast. *Ein Dialogmodul für ein Spracherkennungs- und Dialogsystem*, volume 50 of *Dissertationen zur künstlichen Intelligenz*. infix, St. Augustin, 1993.

[MKE+94] M. Mast, F. Kummert, U. Ehrlich, G. Fink, T. Kuhn, H. Niemann, and G. Sagerer. A Speech Understanding and Dialog System with a Homogeneous Linguistic Knowledge Base. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 16(2):179–16, 1994.

[MKK+92] M. Mast, R. Kompe, F. Kummert, H. Niemann, and E. Nöth. The Dialog Module of the Speech Recognition and Dialog System EVAR. In *Int. Conf. on Spoken Language Processing*, volume 2, pages 1573–1576, Banff, 1992.

[Nöt91] E. Nöth. *Prosodische Information in der automatischen Spracherkennung — Berechnung und Anwendung*. Niemeyer, Tübingen, 1991.

[NSSK90] H. Niemann, G. Sagerer, S. Schröder, and F. Kummert. ERNEST: A Semantic Network System for Pattern Analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 9:883–905, 1990.

[O'S92] D. O'Shaughnessy. Analysis of False Starts in Spontaneous Speech. In *Int. Conf. on Spoken Language Processing*, volume 2, pages 931–934, Banff, 1992.

[OWV93] M. Ostendorf, C.W. Wightman, and N.M. Veilleux. Parse Scoring with Prosodic Information: an Analysis/Synthesis approach. *Computer Speech & Language*, 7(3):193–210, 1993.

[SBD92] E. Shriberg, J. Bear, and J. Dowding. Automatic Detection and Correction of Repairs in Human-Computer Dialog. In *DARPA Speech and Natural Language Workshop*, page 6 Seiten, Arden House, N.Y., 1992.

[SL92] E.E. Shriberg and R.J. Lickley. Intonation of Clause-Internal Filled Pauses. In *Int. Conf. on Spoken Language Processing*, volume 2, pages 991–994, Banff, 1992.

[STNE+93] E.G. Schukat-Talamazzini, H. Niemann, W. Eckert, T. Kuhn, and S. Rieck. Automatic Speech Recognition without Phonemes. In *Proc. European Conf. on Speech Communication and Technology*, volume 1, pages 111–114, Berlin, September 1993.

[Vai88] J. Vaissière. The Use of Prosodic Parameters in Automatic Speech Recognition. In H. Niemann, M. Lang, and G. Sagerer, editors, *Recent Advances in Speech Understanding and Dialog Systems*, volume 46 of *NATO ASI Series F*, pages 71–99. Springer–Verlag, Berlin, Heidelberg, New York, 1988.

[Wai88] A. Waibel. *Prosody and Speech Recognition*. Morgan Kaufmann Publishers Inc., San Mateo, California, 1988.

[WH92] M.Q. Wang and J. Hirschberg. Automatic Classification of Intonational Phrase Boundaries. *Computer Speech & Language*, 6(2):175–196, 1992.