BMFT
gefördert vom
Bundesministerium
für Forschung
und Technologie

Verbmobil
Verbundvorhaben

PROSODY TAKES OVER: TOWARDS A PROSODICALLY GUIDED DIALOG SYSTEM

R. Kompe, E. Nöth, A. Kießling,
T. Kuhn, M. Mast, H. Niemann,
K. Ott, A. Batliner

F.–A.–Universität Erlangen–Nürnberg
L.M.-Universität München

Vm

Report 48
Dezember 1994

Dezember 1994

R. Kompe, E. Nöth, A. Kießling, T. Kuhn, M. Mast, H. Niemann, K. Ott,
A. Batliner

Lehrstuhl für Mustererkennung (Inf. 5)
Friedrich–Alexander–Universität Erlangen–Nürnberg
Martensstr. 3
D–91058 Erlangen

Institut für Deutsche Philologie
Ludwig–Maximilian Universität München
Schellingstr. 3
D–80799 München

Tel.: (09131) 85 - 7890
e-mail: kompe@informatik.uni-erlangen.de

Gehört zum Antragsabschnitt: 3.11, 3.12, 6.4

Das diesem Bericht zugrundeliegende Forschungsvorhaben wurde mit Mitteln des Bundesministers für Forschung und Technologie unter dem Förderkennzeichen 01 IV 102 H/0 und 01 IV 102 C 6 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei dem Autor.

From Speech Communication Vol. 15, Nos. 1-2, October 1994, pp. 155–167.

PROSODY TAKES OVER: TOWARDS A PROSODICALLY GUIDED DIALOG SYSTEM

R. Kompe, E. Nöth, A. Kießling, T. Kuhn, M. Mast,
H. Niemann, K. Ott
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

Address for correspondence:

R. Kompe
Universität Erlangen-Nürnberg
Lehrstuhl fuer Mustererkennung (Informatik 5)
Martensstr. 3
91058 Erlangen
Germany
phone: +49/9131/857890
fax: +49/9131/303811

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 often determined by prosodic cues only. An important result of experiments where naive persons used the EVAR system is that it is hard to follow the train connection given 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 this utterance.

Zusammenfassung. Der Diskursbereich des Spracherkennungs- und Dialogsystems EVAR ist Fahrplanauskunft für Züge. Wir beobachteten, daß in realen Mensch–Mensch Dialogen der Kunde sehr oft den Auskunftsbeamten unterbricht, wenn dieser die Information übermittelt. Viele dieser Unterbrechungen sind ausschließlich Wiederholungen der Uhrzeitangabe des Beamten. Die funktionale Rolle dieser Unterbrechungen wird häufig alleine durch prosodische Mittel bestimmt. Ein wichtiges Ergebnis von Dialog–Experimenten mit naiven Personen ergab, daß es schwer ist, den Verbindungsauskünften von EVAR via Sprachsynthese zu folgen. In diesem Fall ist es sogar noch wichtiger als in Mensch–Mensch Dialogen, daß der Benutzer die Möglichkeit hat, während der Antwortphase zu interagieren. Deshalb haben wir das Dialogmodul erweitert, um dem Benutzer die Möglichkeit zu geben, die Uhrzeitangaben zu wiederholen, und wir fügten ein Prosodiemodul hinzu, das die Fortführung des Dialogs steuert, indem die Intonation dieser Äußerung analysiert wird.

Résumé. Le domaine du système de reconnaissance de la parole et de dialogue EVAR comprend des renseignements d'horaires de train. Nous avons constaté que dans les dialogues réels d'homme à homme, la personne qui cherche une information interrompt souvent l'agent lorsque celui-ci communique l'information. La plupart de ces interruptions sont des répétitions d'horaires indiqués par l'agent. Le rôle fonctionel de ces interruptions est determiné uniquement par des moyens prosodiques. Un resultat essentiel obtenu par une multitude d'expériences effectueés avec des personnes naives est le fait qu'il est difficile de suivre les informations d'horaires d'EVAR par la synthese de la parole. Dans ce cas, il est encore plus important que dans les dialogues réels d'homme à homme que l'usager puisse intervenir lors de la réponse. Voilà pourquoi nous avons élargi le module dialogue pour lui donner la possibilité de répéter les horaires et nous avons de même ajouté un module prosodique commandant la poursuite du dialogue en analysant l'intonation du commentaire.

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 (train time table inquiries) of our speech understanding and dialog system EVAR 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. Waibel, 1988). 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 (1991), 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 a brief 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.

Input to the system is continuous German speech. 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 Schukat–Talamazzini et al., 1993). The lexicon of the system contains 1081 words.

All the linguistic (i.e., syntactic, semantic, and dialog) knowledge is integrated in a homogeneous knowledge base (the semantic network shell ERNEST, see Niemann et al., 1990). This system architecture makes constraint propagation during analysis across all linguistic levels easy. The control algorithm used for the analysis is defined within ERNEST and basically does not depend on the application. It is based on an A^*-search. For a more detailed description of the EVAR system see Mast et al. (1994).

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 relevant results of experiments with the system are presented (for more details see Mast et al., 1992, Mast, 1993).

In a user–friendly system the user should have the possibility of talking to the system without extensive restrictions, i.e., almost as if (s)he were talking to an information officer. The dialog model must therefore represent all expected sequences of dialog acts

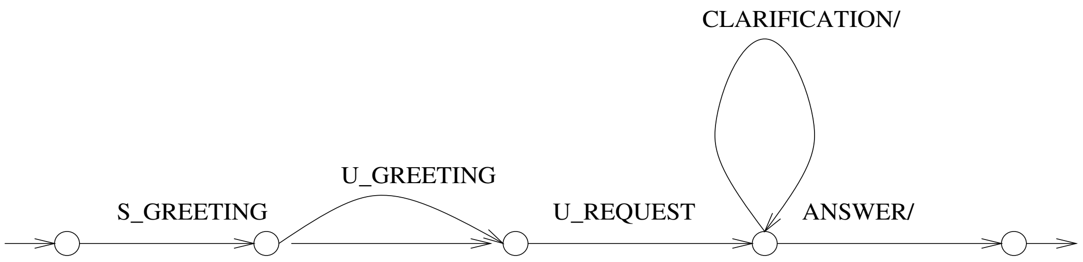

Figure 1: Recursive transition network representing the dialog model implemented in EVAR.

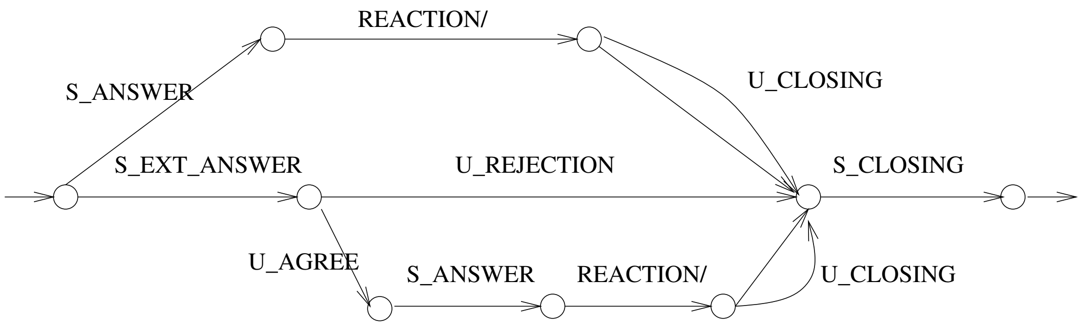

Figure 2: The ANSWER/ subnet.

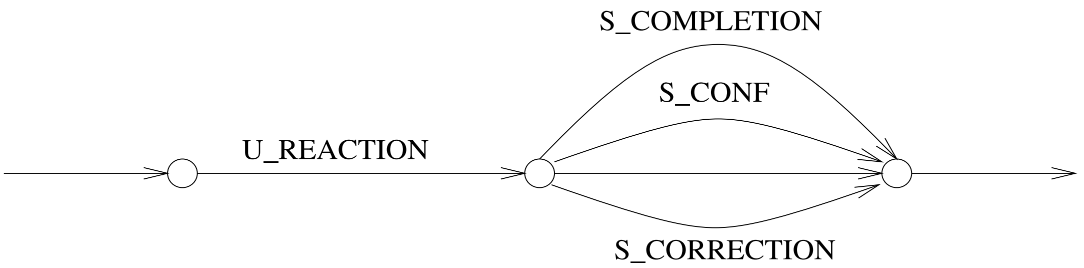

Figure 3: The REACTION/ subnet (cf. section 5.2).

which are typical in this special situation. From a corpus of real human–human dialogs (cf. Hitzenberger et al., 1986), a model was extracted containing all the sequences of dialog acts observed in the corpus which are relevant for human–machine communication. Figure 1 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 2 shows the subnet for the answer phase (ANSWER/). The subnet "REACTION/" (Figure 3) 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, or if part of the user utterance could not be analyzed, 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 actual answers are given via the speech synthesis system SPRAUS from Daimler–Benz, Ulm. The following examples for the different dialog phases are translated into English (the abbreviations of Figures 1 and 2 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.

With this system, experiments with 15 naive subjects were conducted (cf. Mast, 1993, Niemann et al., 1994):

Forty of a total of 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. Many of the misunderstandings were due to spontaneous speech phenomena such as false starts, repetitions, filled pauses and non–speech events (cf. O'Shaughnessy, 1992, Shriberg and Lickley, 1992a, Shriberg et al., 1992b) which are not yet modeled by the word recognizer (compare Butzberger et al., 1992) and not yet taken into account during linguistic analysis. Further, 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 Daly and Zue, 1990, Daly and Zue, 1992, Batliner et al., 1994, Batliner et al., 1993). To assess user satisfaction after each session the users were asked to answer a questionnaire. Twelve of the 15 users suggested a few improvements, especially that the answers should be presented slower, and with a possibility for repetition.

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, 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

confirmation:		
officer: *You'll arrive in Munich at 5 32.*		
customer:	**5 32.**	
question:		
officer: *...you'll leave Hamburg at 10 15...*		*...yes, 10 15, and you'll reach...*
customer:	**10 15 ?**	
feedback:		
officer: *...the next train leaves at 6 35...*		*...and arrives in Berlin at 8 15.*
customer:	**6 35 –**	

Table 1: Examples for officer answer, user interruption, and officer reaction.

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 subsequent action of the officer. In this section we will summarize the main results of this investigation. For further details see Batliner et al. (1992).

4.1 Customer Interruptions: F0–contours and Functional Roles

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. In general, there are two types 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 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* (for examples see Table 1; for corresponding F0–contours see Figures 4–6).

- Using a **confirmation**, the customer wants to signal the officer that he received the last information, e.g., the time of arrival. Functionally, this corresponds to the word *"Roger"* in radio traffic. Usually, the intonation ($F0$–contour) at the end of such an

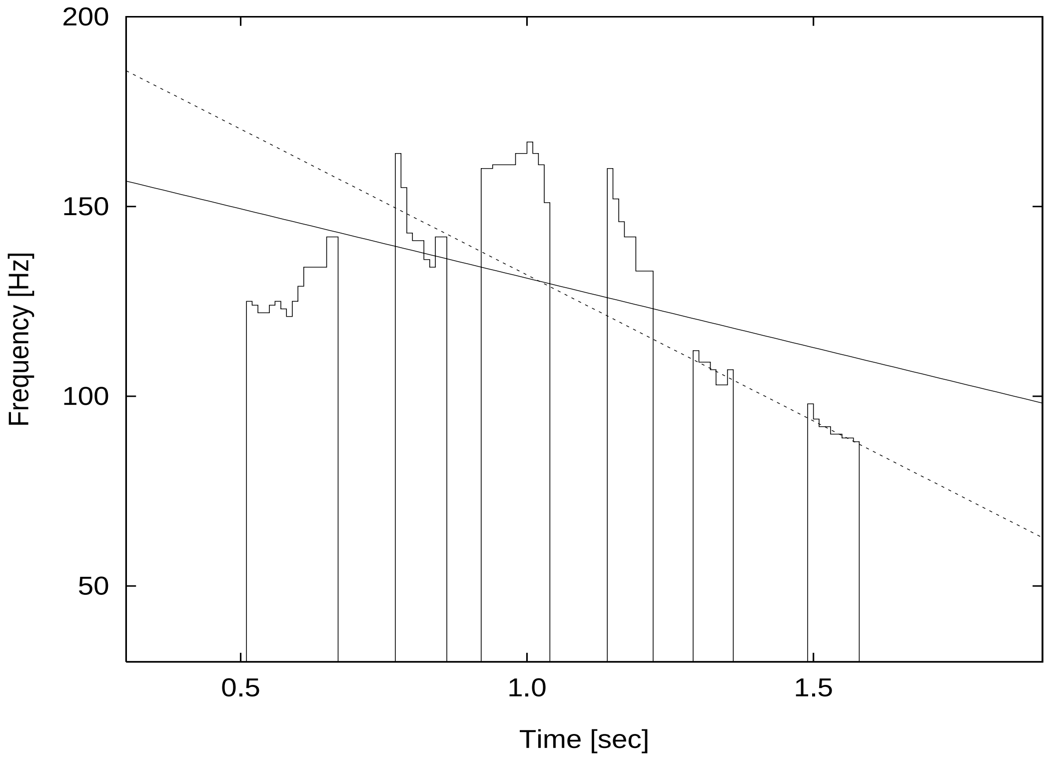

Figure 4: Protoypical *falling* F0-contour and regression line over the whole utterance (solid line) and over the last two voiced regions (dashed line); functional role: *confirmation*.

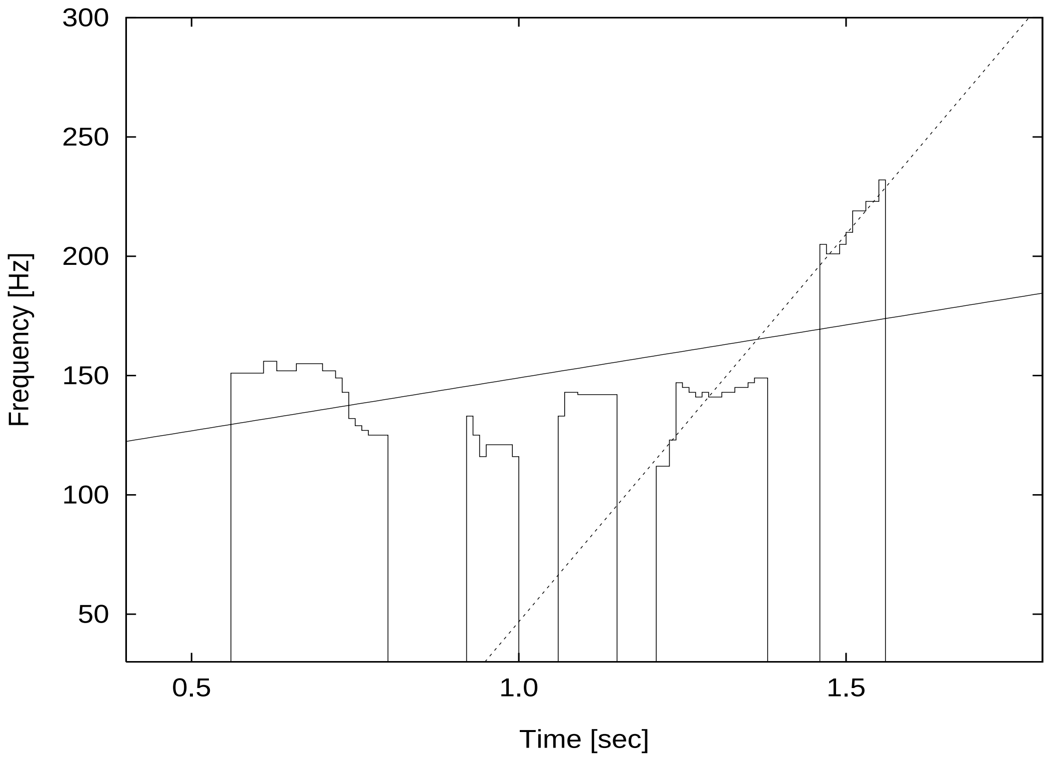

Figure 5: Protoypical *rising* F0-contour and regression line over the whole utterance (solid line) and over the last two voiced regions (dashed line); functional role: *question*.

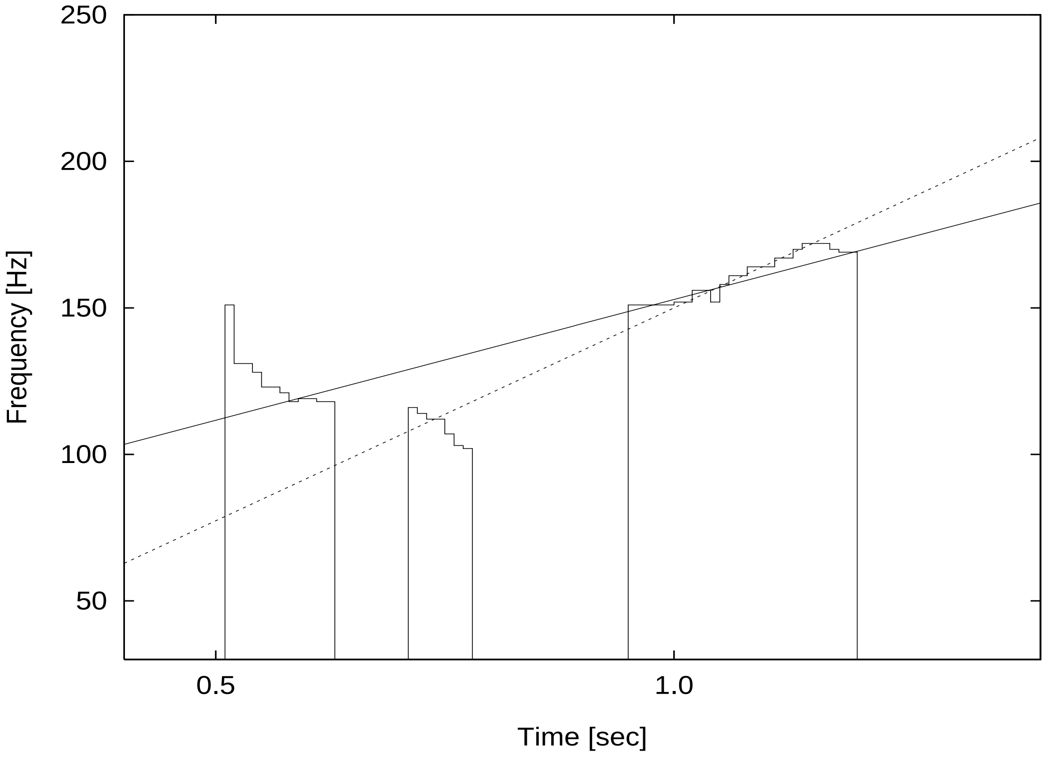

Figure 6: Protoypical F0-contour for *continuation-rise* and regression line over the whole utterance (solid line) and over the last two voiced regions (dashed line); functional role: *feedback*.

utterance is falling (see Figure 4). A confirmation can frequently be observed after the end of a turn of the officer, just at the beginning of the turn–taking by the customer.

- The function of a **question** is *"Sorry, please repeat"*. The customer signals the officer that he did not understand, that he did not get the time–of–day completely or that he just wants to ask the officer to confirm the correctness (*"correct me if I'm wrong"*). The prototypical $F0$–contour is rising (see Figure 5). These questions often occur as short interruptions during the answer phase of the officer.
- By using a **feedback**, the customer usually wants to signal the officer *"I'm still listening"*, *"I got the information"* and sometimes *"slow down, please!"* or *"just let me take down the information"*. It is usually characterized by a level or slightly rising $F0$–contour (continuation rise, see Figure 6) and, like the question, it is usually found during the answer phase of the officer.

Note that one has to distinguish **function** (*confirmation, question*, and *feedback*) and intonational **form** (*fall*, *rise*, and *continuation rise*) although in prototypical cases there is an unequivocal mapping of form onto function. The dialog guiding function of a confirmation is similar to a feedback, but their intonational form is different. Usually, questions can be distinguished easily from confirmations. Feedbacks, however, are sometimes likely to be confused with questions or even with confirmations.

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 on which the functional role (confirmation, etc.) did not depend, such as "*Leave Munich at five seventeen*".

Just as in these human–human dialogs elliptic repetitions of parts of information can often be observed in simulations of human–machine dialogs as well (cf. Krause et al., 1990, Hitzenberger and Kritzenberger, 1989). Therefore we intended to take into account in the dialog model of our system that the continuation of the dialog can be controlled by intonation. To simplify the problem for the beginning we restricted our model to user utterances where only the intonation and no grammatical indicators govern the system reaction. Further, only isolated time–of–day repetitions are considered which are the majority of the 100 cases mentioned above.

4.2 The Scheme of Officer Reactions

From the corpus we developed a scheme (see Table 2) showing the reactions of the officer depending on the intonation of the repetition of time–of–day by the customer. The intonation contour was classified manually by an expert listening to the signals. 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 5.2).

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 2, column 1). Column 2 of Table 2 shows the type of the intonation contour of the customer utterance. The entries in the first two columns completely determine the reaction of the officer (column 3 of Table 2), 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 2 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 5.2). 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 (s)he understood the officer utterance and do not provoke any special reaction by the officer. When the customer repeats the

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 & incom- plete	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 2: The reaction scheme for repetitions of the time of day (RTD) within the dialog system EVAR. (The word "*Uhr*" means "*hour*", "*nein*" = "*no*", "*ja*" = "*yes*", "*um*" = "*at*".)

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.

5 The Dialog Module with Prosody

To cope at least partly with the problems mentioned in Section 3, 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 conducted experiments which led to an automatic classifier of sentence modality (i.e., fall, rise and continuation rise), that are mapped prototypically onto the functional roles of the repetition of time–of–day (i.e., confirmation, question, and feedback).

For training and testing of the classifier, two databases were recorded and digitized with 16 kHz and 14 bits: In database A one female and three male speakers (not "naive", because they are working on prosody) each read the same 90 complete time–of–day utterances (all with the word *"Uhr"*; 30 questions, confirmations, and feedbacks respectively). As this database was used for training, misproductions (e.g., a question was intended, but a falling $F0$–contour was produced) and erroneous $F0$–contours were discarded by visual comparison between the speech signal and the $F0$–contour and by auditory tests. Thus a total of 322 utterances could be used for training. In database B two female and two male "naive" speakers read 50 time–of–day expressions each (47% of them contained the word *"Uhr"*). Neither misproductions nor erroneous $F0$–contours were sorted out; this database, therefore, gives a good impression of how the system could perform in a real environment.

From the automatically computed $F0$–contour (cf. Kießling et al., 1992) a number of features were computed. The best results were obtained using the following four features that were extracted by considering only the voiced frames (non–zero values): the slope of the regression line of the whole (see the solid lines in Figures 4 to 6) and of the last two voiced regions of the $F0$–contour (dashed lines in Figures 4 to 6), 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 (related work and comparable features are, e.g., reported in Waibel, 1988, Daly and Zue, 1990, Daly and Zue, 1992). Gaussian classifiers with full covariance matrix were trained to classify into the three classes fall (F), rise (R), and continuation rise (CR) and thus — prototypically — into the functional roles confirmation, question, and feedback.

Three experiments were performed. In the first experiment, database A was used for testing in a leave–one–out mode (three speakers in turn were used for training, the other for testing). In the second experiment, the classifier trained on database A was tested on database B. Different feature combinations (e.g., computing the slope of the second regression line over the last, the last two or the last three voiced regions) were tried. For the best feature combination where the second regression line was computed over the last two voiced regions, confusion matrices are given in Tables 3 and 4 (rows: spoken classes — number of occurrences in parentheses; columns: recognized classes; numbers are in percent). In the leave–one–out experiment (see Table 3) for all 3 classes, good recognition rates could be achieved (average recognition rate: 87.5%). For the speaker–independent test with the naive speakers (see Table 4) we obtained an average recognition rate of 71.3%. Whereas questions and confirmations were recognized with approximately the same recognition rate (88%) as in the first experiment, it was much more difficult to classify the feedbacks correctly. The reason might be that database B was not controlled with respect to erroneous $F0$-values and — more importantly — with respect to misproductions; it turned out to be difficult for naive speakers to produce a continuation rise correctly while reading an utterance. This is not the case in real–life.

As a final experiment the classifier trained on database A was tested on a subset of the "real–life" material mentioned in Section 4.2. Due to the sometimes very noisy telephone quality, only 32 isolated repetitions of time–of–day could be used for classification. Their reference type (fall, rise or continuation rise) was determined by auditory tests and acoustic measurements. For automatic classification, the same features as described

	R	CR	F
R (97)	81.4	18.6	0.0
CR (107)	7.5	87.9	4.7
F (118)	1.7	5.1	93.2

Table 3: Classification results on database A (leave-one-out training/testing). R: rise, F: fall, CR: continuation-rise.

	R	CR	F
R (70)	87.1	7.1	5.7
CR (64)	21.9	37.5	40.6
F (66)	3.0	7.6	89.4

Table 4: Classification results for database B (training with database A); R: rise, F: fall, CR: continuation-rise.

above were extracted from the digitized signal. and the same classifier was used. All the 10 confirmations, all the 5 questions and 7 of the 17 feedbacks were classified correctly (this is a total recognition rate of 69%).

Note that if a confirmation is misclassified as a question it has no dramatic consequences: the system just gives redundant information the user did not ask for. However, when a question is misclassified as a confirmation, the user does not get the requested repetition of the time–of–day. A confusion of feedback with confirmation in most cases has no effect on the reaction of our system.

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 3). 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 3). 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 2). 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 2). Which one of these alternatives is chosen depends on the reaction scheme of Table 2, 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 linked to the following concepts (cf. also Section 5.4):

- a concept representing silence. During analysis at first it is tried if for this concept an instance can be created, 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. An instance for this concept is created if the creation of an instance of the silence concept failed. This concept itself has 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 5.1. 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. Section 6) 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 infinitive 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 (compare Price et al., 1990).

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 5.2 (subnet REACTION/, see Figure 3). As pointed out in Section 5.2, 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 2). 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 5.2) 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].

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 2, 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 2.

6 Discussion and Future Work

Prosodic information can be used on all levels of speech understanding and dialog. However, few applications have been published: Waibel (1988), Nöth and Kompe (1988) and Hieronymus et al. (1992) use accent information for word recognition; Ostendorf et al. (1993) report on the disambiguation of utterances based on the comparison of alternative parses with information about prosodic phrase boundaries; Robinson et al. (1990) use F0 as an additional feature to enhance phoneme recognition; Singer and Sagayama (1992) use F0 to normalize the spectral features for phone recognition; Kenny et al. (1991) use duration for word recognition.

Already Lea (1980) and Vaissière (1988) discussed the integration of a prosodic module into automatic speech understanding (ASU) systems. 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 still is at an experimental stage, i.e., the user, so far, cannot really interrupt a system utterance, but after each system utterance 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.

Batliner et al. (1993) showed that spontaneous speech contains a lot of elliptic utterances and that, in general, the intonational marking of sentence modality is rather distinct in elliptic utterances. Therefore we expect our modeling of question, confirmation, and feedback with *rise*, *fall*, and *continuation rise* to work reasonably well not only with repetitions of time–of–day in train time table dialogs, but also within other scenarios, where any short elliptic utterances in clarification dialogs are used, e.g., dialogs in which appointments are to be made — this is the scenario in the VERBMOBIL project (automatic translation of face–to–face dialogs) which we are involved in (cf. Wahlster, 1993). However, our modeling is not exhaustive. If, for example, in a confirmation, an (contrastive) accent is positioned on the last syllable, or vice versa, in a question on the first syllable, our model will possibly not work adequately. To cope with this problem either better features have to be found, which take accentuation into account, or the spoken F0 contours have to be matched with prototypical F0–contours using a method such as dynamic time warping. Moreover, repetitions of time–of–day might not be purely isolated. They often do occur together with particles (such as *"yes", "no")* or with repetitions of city names.

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 (Bakenecker et al., 1994, compare Ostendorf et al., 1993) and for re–scoring the n–best sentence hypotheses is being explored. Results concerning the recognition of prosodic phrase boundaries and phrase accents are presented by Kompe et al. (1994a) and Kießling et al. (1994) (see also Wightman et al., 1992, Veilleux et al., 1990, Wang and Hirschberg, 1992 and Huber, 1989). In these contexts not only the prosodic parameter intonation is considered but duration and intensity as well.

Acknowledgements

This work was supported by the German Ministry for Research and Technology (*BMFT*) in the joint research projects ASL and VERBMOBIL and by the German Research Foundation (DFG). Only the authors are responsible for the contents of this paper. We wish to thank S. Hunnicutt and the unkown reviewer for their helpful comments.

References

[Bakenecker et al., 1994] Bakenecker, G., Block, U., Batliner, A., Kompe, R., Nöth, E., and Regel-Brietzmann, P. (1994). Improving Parsing by Incorporating 'Prosodic Clause Boundaries' into a Grammar. In *Int. Conf. on Spoken Language Processing*, volume 3, pages 1115–1118, Yokohama.

[Batliner et al., 1992] Batliner, A., Kießling, A., Kompe, R., Nöth, E., and Raithel, B. (1992). 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.

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

[Batliner et al., 1993] Batliner, A., Weiand, C., Kießling, A., and Nöth, E. (1993). Why sentence modality in spontaneous speech is more difficult to classify and why this fact is not too bad for prosody. In *Proc. ESCA Workshop on prosody*, pages 112–115, Lund.

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

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

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

[Hieronymus et al., 1992] Hieronymus, J., McKelvie, D., and McInnes, F. (1992). Use of Acoustic Sentence Level and Lexical Stress in HSMM Speech Recognition. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 225–228, San Francisco.

[Hitzenberger and Kritzenberger, 1989] Hitzenberger, L. and Kritzenberger, H. (1989). 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.

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

[Huber, 1989] Huber, D. (1989). 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.

[Kenny et al., 1991] Kenny, P., Parthasarathy, S., Gupta, V., Lenning, M., Mermelstein, P., and O'Shaughnessy, D. (1991). Energy, Duration and Markov Models. In *Proc. European Conf. on Speech Communication and Technology*, volume 2, pages 655–658, Genova.

[Kießling et al., 1994] Kießling, A., Kompe, R., Batliner, A., Niemann, H., and Nöth, E. (1994). Automatic Labeling of Phrase Accents in German. In *Int. Conf. on Spoken Language Processing*, volume 1, pages 115–118, Yokohama.

[Kießling et al., 1992] Kießling, A., Kompe, R., Niemann, H., Nöth, E., and Batliner, A. (1992). 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.

[Kompe et al., 1994] Kompe, R., Batliner, A., Kießling, A., Kilian, U., Niemann, H., Nöth, E., and Regel-Brietzmann, P. (1994). 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.

[Krause et al., 1990] Krause, J., Hitzenberger, L., Krischker, S., Kritzenberger, H., Mielke, B., and Womser-Hador, C. (1990). Endbericht zum BMFT-Projekt "Sprachverstehende Systeme; Teilprojekt Simultaion einer multimedialen Dialog-Benutzer Schnittstelle – DICOS". FG Linguistische Informationswissenschaft Universität Regensburg.

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

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

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

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

[Niemann et al., 1994] Niemann, H., Eckert, W., Kießling, A., Kompe, R., Kuhn, T., Nöth, E., Mast, M., Rieck, S., Schukat-Talamazzini, E., and Batliner, A. (1994). Prosodic Dialog Control in EVAR. In Niemann, de Mori, and Hanrieder, editors, *Progress and Prospects of Speech Research and Technology: Proc. of the CRIM/FORWISS Workshop (München, Sept. 1994)*, pages 166–177, Sankt Augustin. infix.

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

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

[Nöth and Kompe, 1988] Nöth, E. and Kompe, R. (1988). Der Einsatz prosodischer Information im Spracherkennungssystem EVAR. In Bunke, H., Kübler, O., and Stucki, P., editors, *Mustererkennung 1988 (10. DAGM Symposium)*, volume 180 of *Informatik FB*, pages 2–9. Springer–Verlag, Berlin.

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

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

[Price et al., 1990] Price, P., Wightman, C., Ostendorf, M., and Bear, J. (1990). The Use of relative Duration in Syntactic Disambiguation. In *Int. Conf. on Spoken Language Processing*, volume 1, pages 13–18, Kobe, Japan.

[Robinson et al., 1990] Robinson, T., Holdsworth, J., Patterson, R., and Fallside, F. (1990). A Comparison of Preprocessors for the Cambridge Recurrent Error Propagation Network Speech System. In *Int. Conf. on Spoken Language Processing*, volume 2, pages 1033–1036, Kobe.

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

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

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

[Singer and Sagayama, 1992] Singer, H. and Sagayama, S. (1992). Pitch Dependent Phone Modelling for HMM Based Speech Recognition. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 273–276, San Francisco.

[Vaissière, 1988] Vaissière, J. (1988). The Use of Prosodic Parameters in Automatic Speech Recognition. In Niemann, H., Lang, M., and Sagerer, G., 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.

[Veilleux et al., 1990] Veilleux, N., Ostendorf, M., Price, P., and Shattuck-Hufnagel, S. (1990). Markov modeling of prosodic phrase structure. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 777–780, Albuquerque.

[Wahlster, 1993] Wahlster, W. (1993). Verbmobil — Translation of Face–To–Face Dialogs. In *Proc. European Conf. on Speech Communication and Technology*, volume "Opening and Plenary Sessions", pages 29–38, Berlin.

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

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

[Wightman et al., 1992] Wightman, C., Shattuck-Hufnagel, S., Ostendorf, M., and Price, P. (1992). Segmental durations in the vicinity of prosodic boundaries. *J. of the Acoustic Society of America*, 91:1707–1717.