gefördert vom
Bundesministerium
für Forschung
und Technologie

Verbmobil
Verbundvorhaben

PROSODY TAKES OVER: A PROSODICALLY GUIDED DIALOG SYSTEM

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

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

Vm Report 47
Dezember 1994

Dezember 1994

R. Kompe, A. Kießling,
T. Kuhn, M. Mast,
H. Niemann, E. Nöth,
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 Proc. EUROSPEECH 1993, Berlin, Vol. 3, pp. 2003–2006

PROSODY TAKES OVER: A PROSODICALLY GUIDED DIALOG SYSTEM

R. Kompe[1], **A. Kießling**[1], **T. Kuhn**[1], **M. Mast**[1], **H. Niemann**[1], **E. Nöth**[1], **K. Ott**[1], **A. Batliner**[2]
[1]Universität Erlangen-Nürnberg, Lehrstuhl für Mustererkennung (Informatik 5), Martensstr. 3, 91058 Erlangen, FRG
e-mail: kompe@informatik.uni-erlangen.de
[2]L.M.-Universität München, Institut für Deutsche Philologie, Schellingstr. 3, 80799 München, FRG

Abstract: *In this paper first experiments with naive persons using the speech understanding and dialog system EVAR are discussed. The domain of EVAR is train 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 the experiments with EVAR is that it is hard to follow the 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.*

Keywords: ASU, prosody, dialog

1. Introduction

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 the German *InterCity* train system.

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 [10]). The lexicon of the system contains 1081 words.

All the linguistic knowledge is integrated in a homogeneous knowledge base (the semantic network shell ERNEST, see [9]). 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 the A^*−Algorithm. For a more detailed description of the EVAR system see [8].

The paper is organized as follows: First we give an overview about the dialog module without prosody (for details see [7, 6]). Then results of recent experiments with naive persons using EVAR are presented. Motivated by these and by the observation of real human-human dialogs (section 3) we extended the dialog module and added a prosody module to the system, which is described in the final part of the paper (section 4).

2. 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 also the consideration of the dialog history in order to be able to resolve references and to focus the analysis on the expected answer.

In a user-friendly system the user should have the possibility to talk to the system without many restrictions, i.e., almost like talking to an information officer. So the dialog model must represent all dialog acts which are typical in this special situation. From a corpus of real human-human dialogs [2] a model was extracted containing all sequences of dialog acts observed in the corpus. We achieved a simplification compared to real natural dialogs by guiding the user with special system utterances. Figures 1 to 3 show part of 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 system or a user utterance, respectively. The subnet for clarification will not be discussed in this paper. Figure 2 shows the subnet for the answer phase. The subnet "REACTION/" (figure 3) contains the extensions to the dialog model relevant for this paper. It is described in section 4 and was not implemented in the version of the system that was used for the experiments presented 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 for information. If the information that is necessary for giving an answer is not given in the user's request the system starts a clarification dialog (which is not topic of this paper). 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 figure 1 and 2 are used):

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

Request:
U: (U_REQ_INFO) I want to go to Hamburg tomorrow in

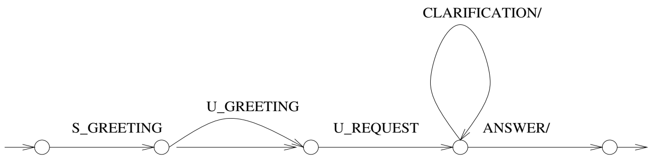

Figure 1: The dialog model implemented in EVAR.

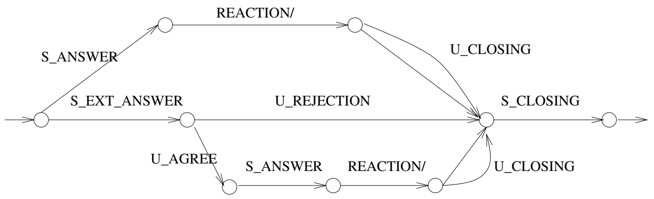

Figure 2: The ANSWER/ subnet.

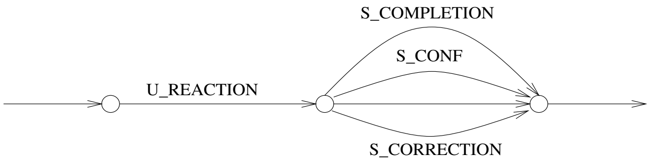

Figure 3: The REACTION/ subnet (cf. section 4).

the afternoon.

Answer:
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.

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

With this system experiments with 15 naive subjects were conducted:

82 dialogs were recorded by naive users. They were asked to take the part of the customer in four different scenarios. Two scenarios were given and equal for every user and the other two scenarios were created by the users themselves. The experiments were conducted in a quiet office environment using a headphone. This was the first experiment with the acoustic module trained on read speech and tested on spontaneous speech.

40 of the 82 dialogs were completed successfully, i.e. the system provided the correct train connection. 8 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 was not completed due to memory limitations, repeated misunderstandings of utterances or if the user gave up the dialog.

The acoustic front-end was trained on 7900 domain specific read sentences from 79 speakers (100 each). A bigram model of perplexity 111 was used. The word accuracy was 73.7% (79.9% of the words and 38.2% of the sentences were correct). For comparison: on read sentences a word accuracy of 92% was achieved.

For the experiments the EVAR system was run on a DECstation 5000/200. The time for the generation of the word hypotheses was 4.2 times realtime. The average CPU time for the linguistic analysis and interpretation in the dialog context for one utterance was 44 sec. The average time to complete a dialog was 9:30 minutes.

In our application the most convenient way to generate an answer 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 as in ours the answer can be quite lengthy. 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 for sure 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 a part of information. As will be seen in the following section this is also the usual way in human-human dialogs.

3. Dialog Guiding Prosodic Signals

We investigated a corpus of 107 "real-life" information retrieval dialogs. The callers did not know that they were recorded. In this section we will summarize the main results of this investigation, for further details see [3]. Subject of the dialogs was train table inquiry. The most important question in this context is how often and in which way during answer generation does the prosody of a user interruption alone control the following actions of the officer.

Just looking at user reactions that contain a repetition of the time of day given by the officer, we observed that in our dialogs on the average they occurred twice per dialog. In half of these cases additional words like in *Excuse me at five seventeen?* indicated the desired action or incorrect repetitions made a correction by the officer necessary regardless of the users intonation. About once per dialog only the intonation was responsible for the appropriate response of the officer. Two thirds of these cases were isolated repetitions of the time of day. The other third contained words that didn't indicate the desired response like in "*Leave Munich at five seventeen*".

We observed three (traditional) categories of $F0$-contours: falling (terminal), rising (interrogative), and slightly rising (continuation rise). A terminal contour signals "roger", i.e., the customer confirms that he understood the time of day. An interrogative contour can be interpreted as "sorry, please repeat". A continuation rise indicates, that the customer is still listening, probably that he takes down the information given by the officer.

Elliptic repetitions of parts of information can often be observed in simulations of human-machine dialogs [4] and a user friendly system should therefore be able to cope with this user behavior. In our system we decided first to model the most frequent and difficult case: the isolated repetition of the time of day. Luckily it turns out that the most difficult case from a "system point of view" is prosodically marked more distinct than the general case, because isolated time of day expressions are elliptic utter-

System answer: *"... In München sind Sie dann um 17 Uhr 32."*
"... You'll arrive in Munich at 5 32 p.m."

<table>
<tr><th colspan="2">RTD</th><th colspan="2">prosody–module</th><th colspan="2">system reaction</th></tr>
<tr><td colspan="2">no utterance</td><td colspan="2">——</td><td colspan="2">——</td></tr>
<tr><td colspan="2">wrong repetition</td><td colspan="2">——</td><td>correction</td><td>('Nein, um 17 Uhr 32.')</td></tr>
<tr><td colspan="2" rowspan="3">complete & correct</td><td>interrogative</td><td>('17 Uhr 32?')</td><td>confirmation</td><td>('Ja, um 17 Uhr 32.')</td></tr>
<tr><td>continuation rise</td><td>('17 Uhr 32–')</td><td colspan="2" rowspan="2">——</td></tr>
<tr><td>terminal</td><td>('17 Uhr 32.')</td></tr>
<tr><td rowspan="6">correct & incom- plete</td><td rowspan="3">only minutes</td><td>interrogative</td><td>('32?')</td><td>confirmation</td><td>('Ja, um 17 Uhr 32.')</td></tr>
<tr><td>continuation rise</td><td>('32–')</td><td colspan="2" rowspan="2">——</td></tr>
<tr><td>terminal</td><td>('32.')</td></tr>
<tr><td rowspan="3">only hours</td><td>interrogative</td><td>('17 Uhr?')</td><td rowspan="2">completion</td><td rowspan="2">('17 Uhr 32.')</td></tr>
<tr><td>continuation rise</td><td>('17 Uhr–')</td></tr>
<tr><td>terminal</td><td>('17 Uhr.')</td><td colspan="2">——</td></tr>
</table>

Table 1: The reaction scheme for repetitions of the time of day (RTD) within the dialog system EVAR. (The word "*Uhr*" means "*hour*".)

ances where the sentence modality cannot be derived from other grammatical indicators like word order or *Wh*-words [1].

4. The Dialog Module with Prosody

To cope at least partly with the problems mentioned at the end of section 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 3 are modeled.

In order to model the potential user reactions we have conducted a couple of experiments which led to an automatic classifier of sentence modality. In such a system it is necessary to determine automatically the category of a $F0$-contour. We therefore recorded a corpus of 360 isolated time of day utterances read by four non-naive speakers (3 male, 1 female, 30 utterances per category and speaker). For each utterance the $F0$-contour was computed automatically. 15 utterances were discarded because of gross $F0$-errors, 23 because of misproduction of the intonation according to perception tests. From the $F0$-contour the following features were extracted: slope of the regression line of the whole and of the final part of the $F0$-contour, and the differences between the offset of the $F0$-contour and the values of the regression lines at the position of the offset. Leave-one-out classification experiments were performed using three speakers for training and one for testing. A Gaussian classifier with full covariance matrix was used. We obtained an average classification rate of 88%. Using all of these utterances for training and 200 read utterances of four other (naive) speakers for testing yielded a recognition rate of 71%. The decrease in performance is due to the fact that no utterances were discarded and that the naive speakers obviously had enormous difficulties in the controlled production of a continuation rise. The recognition rate of interrogative and terminal contours is still at about 88%. On a small set of time of day expressions from the 107 "real-life" dialogs all the 5 interrogative, all the 7 terminal and 7 of the 17 continuation rise $F0$-contours were classified correctly with the same classifier. A more detailed analysis can be found in [3].

In the following we will sketch the analysis process within EVAR after a user utterance has been recorded: The word recognizer computes the best word chain. Since the word recognizer is integrated via procedure call we could 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 the above mentioned corpus of 200 read time of day utterances from 4 speakers is about 82%. Despite the reduced vocabulary and perplexity compared to the results mentioned above the accuracy is lower, because the similarity between the allowed words is way higher. Now the best word chain is semantically interpreted as a time of day expression. 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 system corrects the user and repeats the last answer. In the fourth case the prosody module classifies the intonation contour of the utterance into one of the three classes mentioned above (terminal, continuation rise, interrogative) and the system corrects, completes, confirms or just continues according to the situation. The fifth case can be treated like a complete and correct repetition, i.e. like the fourth case. In the sixth case the system completes the time expression except in the case of a terminal intonation contour. The system corrects, completes, confirms or just continues according to the situation (see figure 3). Table 1 summarizes this system behavior depending on the results of the word recognizer and the prosody module. This scheme was derived from the investigations described in section 3.

In the current system the classification of the intonation contour is done with the Gaussian classifier described above. Implemented is also an alternative approach comparing the actual intonation contour with a set of prototypical F0-contours via dynamic programming (DP). 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.

The prosody module integrated in the semantic network comprises a set of concepts and attributes defining knowledge about the intonation of time of day utterances, performing the classification, and establishing 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 a prosodically guided dialog control. The links to the syntax module were necessary since in the case of classification via DP match, the prosody module has to have access to the word chain underlying the semantic interpretation, and prototypes have to be chosen depending on the number of syllables in the spoken (recognized) word chain.

5. Discussion and Future Work

Already the work of Lea [5] and [11] discuss the integration of a prosodic module into automatic speech understanding (ASU) systems. Lea even proposed a control module very much driven by prosody. However to our knowledge in this paper the first dialog system guided partly by prosodic information is presented. The system still is at experimental stage, i.e. the user so far can not 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 on how to integrate a prosody module in the overall system and getting it to interact with the other system components especially to guide the progress of the dialog.

In the future we plan to work on the integration of prosody on all levels of our ASU system. The integration of accent information into a word recognition module is under investigation. Furthermore the use of prosodic phrase boundaries during syntactic parsing is explored.

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

References

[1] A. Batliner, C. Weiand, A. Kießling, and E. Nöth. *Why sentence modality in spontaneous speech is more difficult to classify and why this fact is not too bad for prosody.* ESCA Workshop on prosody, Lund (Sweden), Sept. 1993.

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

[3] A. Kießling, R. Kompe, H. Niemann, E. Nöth, and A. Batliner. *"Roger", "Sorry", "I'm still listening": Dialog guiding signals in information retrieval dialogs* . ESCA Workshop on prosody, Lund (Sweden), Sept. 1993.

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

[5] 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.

[6] M. Mast. Ein Dialogmodul für ein Spracherkennungs- und Dialogsystem. Dissertation. Technische Fakultät der Universität Erlangen-Nürnberg, 1993.

[7] 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*, pages 1573–1576, Banff, Canada, 1992.

[8] H. Niemann, A. Brietzmann, U. Ehrlich, S. Posch, P. Regel, G. Sagerer, R. Salzbrunn, and G. Schukat-Talamazzini. A Knowledge Based Speech Understanding System. *Int. J. of Pattern Recognition and Artificial Intelligence*, 2(2):321–350, 1988.

[9] 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.

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

[11] 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, 1988.