In Julita Vassileva, Piotr Gmytrasiewicz, and Mathias Bauer (Eds.) (2001). *UM2001, User Modeling: Proceedings of the Eighth International Conference.* Berlin: Springer.

Recognizing Time Pressure and Cognitive Load on the Basis of Speech: An Experimental Study

Christian Müller[1], Barbara Großmann-Hutter[2], Anthony Jameson[2], Ralf Rummer[3], and Frank Wittig[2] *

[1] Department of Computational Linguistics, Saarland University
[2] Department of Computer Science, Saarland University
[3] Department of Psychology, Saarland University

Abstract. In an experimental environment, we simulated the situation of a user who gives speech input to a system while walking through an airport. The time pressure on the subjects and the requirement to navigate while speaking were manipulated orthogonally. Each of the 32 subjects generated 80 utterances, which were coded semi-automatically with respect to a wide range of features, such as filled pauses. The experiment yielded new results concerning the effects of time pressure and cognitive load on speech. To see whether a system can automatically identify these conditions on the basis of speech input, we had this task performed for each subject by a Bayesian network that had been learned on the basis of the experimental data for the other subjects. The results shed light on the conditions that determine the accuracy of such recognition.

1 Background and Issues

This paper is an experimental follow-up to the UM99 paper by Berthold and Jameson ([2]). Those authors argued the following points, among others:

- In a world of increasingly mobile and ubiquitous computing, it is becoming more important for a system ($\mathcal{S}$) to be able to recognize the situation-dependent *resource limitations* of its user ($\mathcal{U}$)—for example, so as to be able to switch to a slower but less demanding style of communication where appropriate (cf. Jameson et al., [5]). While they focused on the variable of cognitive load, we will also consider the variable of time pressure.
- With systems that allow speech input, one source of information about $\mathcal{U}$'s resource limitations is the features of $\mathcal{U}$'s speech; many previous studies have revealed systematic influences of cognitive load (and to a lesser degree, time pressure) on specific features of speech.

On the basis of a synthesis of previous results, Berthold and Jameson ([2]) presented simulations that suggested that it might indeed be feasible to recognize a user's current cognitive load on the basis of a limited amount of speech input; but they noted that

* The research described here was supported by the German Science Foundation (DFG) in its Collaborative Research Center on Resource-Adaptive Cognitive Processes, SFB 378, Projects B2 (READY) and A2 (VEVIAG). We thank Tore Knabe for contributions to the statistical data analyses and Sylvia Bach for help in conducting the experiment.

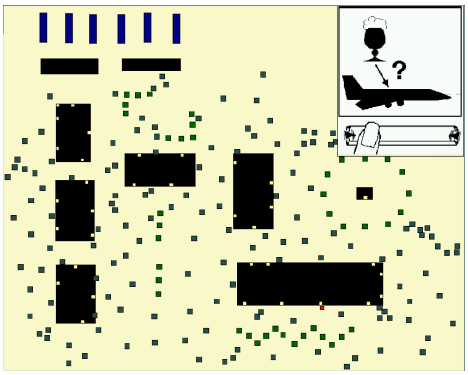

Fig. 1. *Environment used in the experiment, with a typical pictorial stimulus.*

specifically designed empirical studies would be required for a more definite answer to this question.

In the present paper, we first describe an experiment that was explicitly designed to fill this gap (Section 2). We then describe the learning of user models (in the form of Bayesian networks) on the basis of the data from this experiment (Section 3). Finally, we show how well the learned models succeed at recognizing subjects' resource limitations in the experimental data (Section 4).

2 Experiment

2.1 Method

Materials. The experimental environment simulated a situation in which a user is navigating through a crowded airport terminal while asking questions to a mobile assistance system via speech (see Figure 1). In each of 80 trials, a picture appeared in the upper right-hand corner of the screen. On the basis of each picture, the subject was to introduce and ask a question (e.g., "I'm getting thirsty. Is there . . . will it be possible to get a beer on the plane?").

Design. Two independent variables were manipulated orthogonally:

- NAVIGATION? Whether or not the subject was required to move an icon on the screen through the depicted terminal to an assigned destination by pressing arrow keys, while avoiding obstacles and remembering a gate number that comprised five digits and one letter. When navigation was not required, the subject could ignore the depicted terminal and concentrate on the generation of appropriate utterances in response to the pictures.
- TIME PRESSURE? Whether the subject was induced by instructions and rewards (a) to finish each utterance as quickly as possible or (b) to create an especially clear and comprehensible utterance, without regard to time.

Procedure. After an extensive introduction to the scenario, the environment, and the 4 (2×2) conditions, each subject dealt with 4 blocks, each of which comprised

20 stimuli distributed over 4 destinations. Each block was presented in one of the 4 conditions, the order being varied across subjects according to standard procedures.

Subjects. The 32 subjects, students at Saarland University, were paid for their participation. An extra reward was given to one of the participants who most successfully followed the instructions regarding the time pressure manipulation.

Coding and rating of speech. The first author transliterated the subjects' speech input and coded it with respect to a wide range of features, including almost all of those that had been included in previous published studies. On the basis of the transliterations (minus the coding symbols), four independent raters rated the relative "quality" of the 32 utterances produced for each stimulus picture (quality being defined in terms of grammaticality, relevance, clarity, and politeness). The raters also rated the pictorial stimuli in terms of the complexity of the responses that they tended to call for.

In this paper, we report results only for a representative subset of five speech-related variables, which we call *symptoms* because they reflect (albeit imperfectly) the psychological state of the subject induced by the experimental manipulations:

- DISFLUENCIES: The logical disjunction of several binary variables, each of which indexes one feature of speech that involves its formal quality: self-corrections involving either syntax or content; false starts; or interrupting speech in the middle of a sentence or a word.[1]
- ARTICULATION RATE: The number of syllables articulated per second of speaking time, after elimination of the time for measurable silent pauses.
- CONTENT QUALITY: The average quality rank—between 1 (worst) and 32 (best)—assigned to the utterance by the four raters.
- NUMBER OF SYLLABLES: The number of syllables in the utterance.
- SILENT PAUSES: The total duration of the silent pauses in the utterance, expressed relative to the length of the utterance in words (to take into account the fact that longer utterances offer more opportunities for pauses). A silent pause is any silence within the utterance that lasts at least 200 ms.
- FILLED PAUSES: The corresponding measure for filled pauses (e.g., "Uhh").

2.2 Results

Figure 2 shows, for each of the six dependent variables listed above, how it was influenced by the two independent variables TIME PRESSURE? and NAVIGATION?.

DISFLUENCIES. The disfluencies summarized by this variable increased to a significant extent when the speaker was distracted by a navigation task ($F(1,31) = 4.554, p < 0.05$).[2] Perhaps more surprisingly, they increased to almost the same extent when the speaker was *not* under time pressure ($F(1,31) = 4.086$, $p = .052$). The reason may be that subjects in this condition tended to produce longer, more complex utterances (cf. the results for NUMBER OF SYLLABLES shown below), a tendency which is generally associated with a higher frequency of disfluencies (see, e.g., Oviatt, [7]).

[1] Filled and silent pauses are not counted here, because they are treated as separate variables.

[2] The statistical significance of the each of the effects to be discussed in this section was determined through a univariate analysis of variance (ANOVA), in each case after a multivariate ANOVA had shown that the univariate ANOVA was justified.

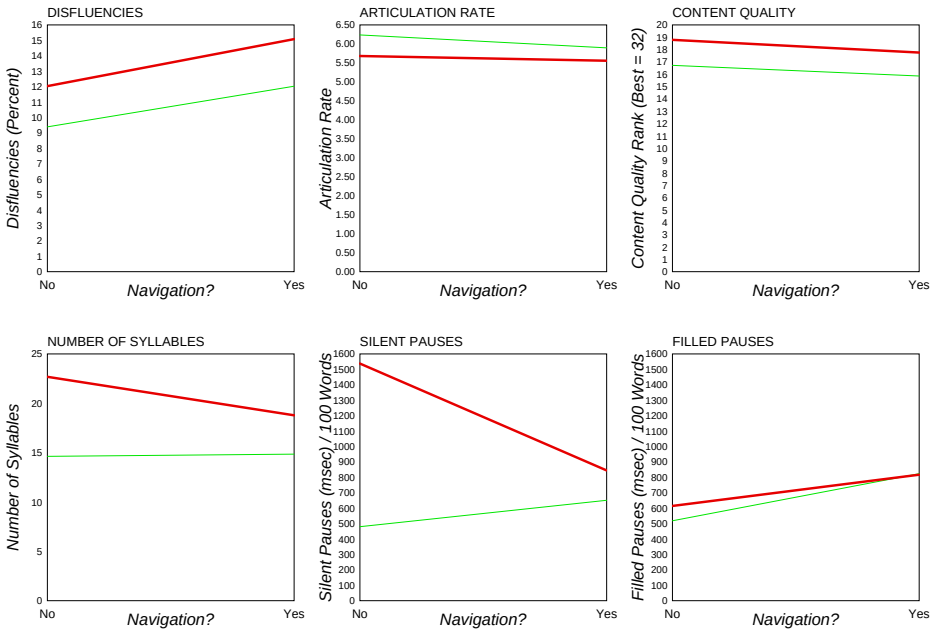

Fig. 2. *Mean values of the six speech symptoms for each of the four experimental conditions.*

(Thin lines: time pressure; thick lines: no time pressure.)

ARTICULATION RATE. On the average, subjects produced more syllables per second when they were under time pressure than when they were not ($F(1,31) = 73.945, p < 0.001$). Though this result is intuitively plausible, it is not logically necessary, given that there are many other ways of coping with time pressure. There is also a statistically highly reliable tendency to articulate less quickly when navigating (see the slope of the two lines; $F(1,31) = 19.958, p < 0.001$), as has been reported in a number of previous studies (cf. Berthold & Jameson, [2]). This effect is stronger under time pressure (interaction: $F(1,31) = 4.297, p < 0.05$).

CONTENT QUALITY. On the average, an utterance produced under time pressure ranks about 2 positions (out of 32) lower than one produced without time pressure ($F(1,31) = 54.986, p < 0.001$). The effect of having to navigate is even smaller, amounting to only about one rank position ($F(1,31) = 12.562, p < 0.001$).

NUMBER OF SYLLABLES. Although the subjects' attempts to produce higher-quality utterances in the absence of time pressure did not lead to much higher quality ratings, they did produce much longer utterances ($F(1,31) = 49.236, p < 0.001$). The most important impact of the navigation task was to reduce this tendency: The increase in length is about 50% without navigation and about 30% with navigation (interaction: $F(1,31) = 9.117, p < 0.01$). Evidently, when they had to navigate, subjects were less ambitious with regard to the goal of producing unambiguous, high-quality utterances.

SILENT PAUSES. The pattern just discussed is even more pronounced for the symptom of silent pauses: The sharp downward slope in the upper line of the graph shows that, when subjects had to navigate, they largely abandoned the goal of generating high-quality utterances that would require careful thought. This effect is especially striking when one considers that a secondary task would in itself tend to increase the number and/or length of silent pauses by demanding the subjects' attention at least intermittently—an effect which is in fact found in the condition with time pressure (lower line) and in previous studies (cf. Berthold, [1]). In sum, the presence or absence of time pressure makes a big difference with regard to silent pauses overall ($F(1,31) = 20.844, p < 0.001$), and the main impact of the navigation task is to reduce this difference (interaction: $F(1,31) = 15.032, p < 0.001$).

FILLED PAUSES. The last graph in Figure 2 shows that filled pauses behave very similarly to silent pauses in the two conditions with time pressure; in particular, they increase when there is a navigation task (as has been shown in previous studies; cf. Berthold, [1]). But in contrast to the case with silent pauses, they show a similar pattern when there is no time pressure. Overall, there is a significant effect of the navigation task ($F(1,31) = 5.924, p < 0.05$) but no significant effect of time pressure and no interaction. In sum, filled pauses might serve as a fairly straightforward index of the presence of a distracting secondary task.

We have seen that each of the dependent variables discussed here shows one or two statistically reliable effects of time pressure and/or the navigation task. These results suggest that observation of these variables in a person's speech might allow a system to infer their current resource limitations. But it is not obvious how successful such diagnosis will actually be in practice. This question is addressed in the next section.

3 Modeling

If we want to create a system that recognizes the resource limitations of its users on the basis of their speech, we need to take two basic steps:

1. Use machine learning methods to create some sort of model relating resource limitations to speech symptoms, using data such as those of this experiment (Sections 3.1 and 3.2 below).
2. Employ this model during an interaction with each user, using the features of their speech as evidence (Section 4).

3.1 Bayesian Network Structure

Regarding Step 1: Among the various techniques that could potentially be used, we employ Bayesian networks (BNs).[3] Within this framework, there are various ways of (a) learning a general user model on the basis of data from a sample of users and (b) adapting it to each individual user on the basis of data about that user (see Jameson & Wittig, [6]). The method employed in the present study is illustrated in Figure 3.

[3] Accessible introductions to BNs are now available from many sources; the classic exposition is that of Pearl ([8]).

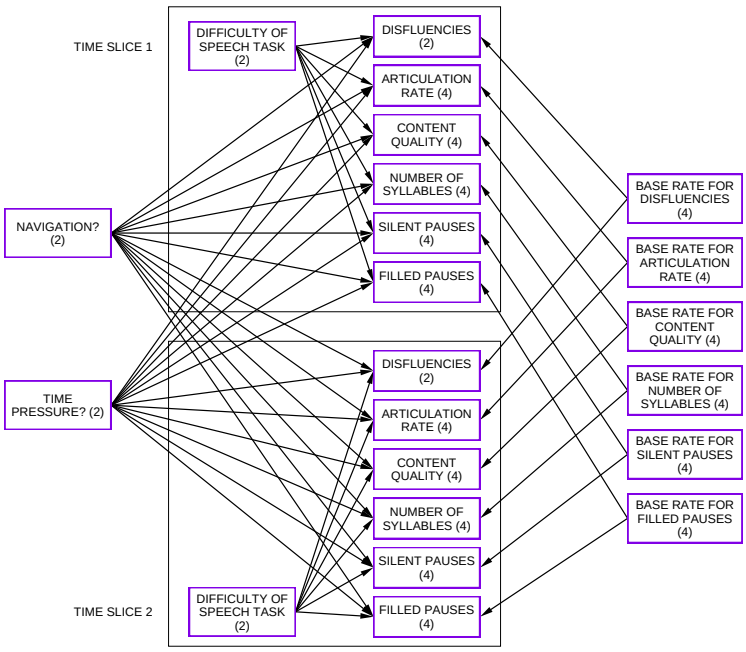

Fig. 3. *Structure of the dynamic Bayesian network used in the evaluation.*
(Nodes within the boxes correspond to temporary variables that index features of the current utterance. Each number in parentheses shows the number of discrete states for the variable in question.)

(The lower part labeled TIME SLICE 2 can be ignored for the moment.) The two nodes NAVIGATION? and TIME PRESSURE? on the left correspond to the two main independent variables of the experiment. The six nodes on the right in TIME SLICE 1 correspond to the dependent variables that we have discussed above.

The six nodes on the far right correspond to individual base rates for the six symptom variables. They are introduced to take into account individual differences in the overall level of the symptom variables. The value of each such variable is constant for each $\mathcal{U}$: It is simply computed as the mean value of the variable in question for the entire experiment.[4]

[4] The BN structure in the figure implies that these base rate variables are statistically independent. This assumption was shown to be false by both structure learning algorithms and factor analyses. Nonetheless, this simplified model was found to perform better at the task of recognizing a speaker's time pressure and cognitive load than did more complex models that took into account the statistical dependencies. A possible reason is that in the more complex models the estimates of some probabilities in the learned BN are less accurate because they are based on relatively few observations. In any case, this result illustrates the general point (see, e.g., Greiner et al., [4]) that the goal in learning BNs is often to learn not the one "correct" model but rather the model which works best for a particular task in a particular setting.

The final node in the BN, DIFFICULTY OF SPEECH TASK, refers to the rated complexity of the speech task created by the stimulus picture (cf. Section 2.1).

Our question in the evaluation study will be: If a user $\mathcal{U}$ produces a sequence of utterances in a given experimental condition, how well can a system $\mathcal{S}$ recognize that condition? Therefore, the variables NAVIGATION? and TIME PRESSURE? can be viewed here as *static* variables whose value does not change over time. The six base rate variables are also static. By contrast, each of the variables inside the boxes labeled TIME SLICE 1 and TIME SLICE 2 refers to an aspect of just one utterance. Hence corresponding *temporary* nodes need to be created for each utterance. We are therefore dealing with a *dynamic Bayesian network* (DBN) that comprises a series of *time slices*.[5]

3.2 Learning of a BN

Since we want to test a learned BN model with the data of a given user $\mathcal{U}$, we must not include $\mathcal{U}$'s data in the data that are used for the learning of the corresponding BN. Accordingly, we learned for each $\mathcal{U}$ the conditional probability tables (CPTs) for a separate BN using the data from the other 31 subjects. The learned BN has the structure shown in Figure 3 minus the nodes shown for TIME SLICE 2; the CPTs for the temporary variables within each time slice are the same as the ones learned for TIME SLICE 1.

The learning method we employed is the usual maximum-likelihood method for learning fully observable Bayesian networks (see, e.g., Buntine, [3]): The estimate of each (conditional) probability is computed simply in terms of the (relative) frequencies in the data.

4 Evaluation

The procedure for evaluating a learned BN is given in Table 1.

Figure 4 shows the results of the evaluation, aggregated over all 32 subjects.[6]

Looking first at the results for recognizing time pressure (left-hand graph), we see that the BNs are on the whole rather successful: The average probability assigned to the actual current condition rises sharply during the first few observations. Note that recognition of time pressure is easier when there is no navigation task.[7] This result is understandable given the overall effects shown in Figure 2: On the whole the effects of time pressure were greatest when there was no navigation task, since speakers could respond more sensitively to the time pressure (or lack of it).

[5] The general principles of dynamic Bayesian networks are explained, e.g., by Russell and Norvig ([9, chap. 17]). A discussion with regard to user modeling of the sort done here is given by Schäfer and Weyrath ([10]). A detailed understanding of DBNs is not required for the reading of this paper.

[6] The results for individual subjects are much less smooth than these aggregated results: The individual curves often show sharp jumps and extreme values.

[7] $p < .01$ for the difference between the average of the two upper curves and the average of the two lower curves. All statistical tests in this section are two-tailed sign tests based on the last 10 observations.

Relevant variables and their values

- A user $\mathcal{U}$
- Values *t* and *n* of the Boolean variables *T* (time pressure?) and *N* (navigation?)

Task

Infer the values of *T* and *N* on the basis of symptoms in $\mathcal{U}$'s speech

Preparation of the test data

Select the 20 observations for $\mathcal{U}$ in which $T = t$ and $N = n$, in the order in which they occurred in the experiment

Evaluating recognition accuracy

Initialize the model:

1. Create the first time slice of the BN for $\mathcal{U}$
2. Instantiate each of the individual baseline variables with its true value for $\mathcal{U}$ (but leave the variables *T* and *N* uninstantiated)

For each observation *O* in the set of observations for $\mathcal{U}$:

1. In the newest time slice of the BN, derive beliefs about *T* and *N*:
   - Instantiate all of the temporary variables for this time slice with their values in *O*
   - Evaluate the BN to arrive at beliefs regarding *T* and *N*
   - Note the probabilities assigned at this point to the true values of *T* and *N*, respectively
2. Add a new time slice to the dynamic BN to prepare for the next observation

Table 1. *Procedure used in evaluating the recognition accuracy of the learned Bayesian networks.*

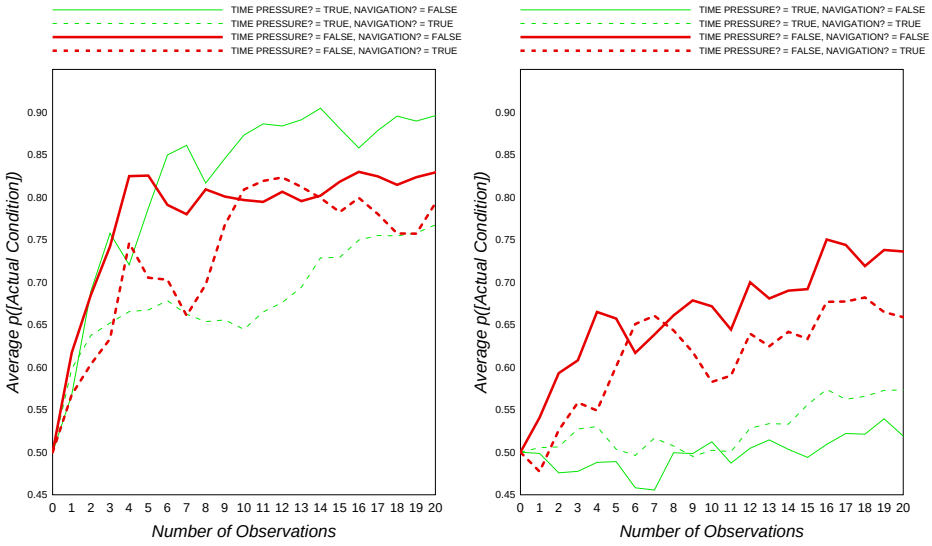

Fig. 4. *Accuracy of the learned BNs in inferring the correct value of* TIME PRESSURE? *(left) and* NAVIGATION? *(right).*

(Each curve shows the aggregated results for one combination of values of the variables TIME PRESSURE? *and* NAVIGATION?*. In each curve, the point for the* i*th observation shows the average probability which the BN assigned to the correct value of the variable in question after processing the first* i *observations.)*

Recognition of the navigation task is considerably less successful: The highest curve in the right-hand graph is significantly lower than the lowest curve in the left-hand graph during the last 10 observations ($p < .01$). This result is likewise understandable given the overall effects shown in Figure 2, where on the whole the effects of time pressure (reflected in the differences between the two lines in each graph) were more substantial than those of the navigation task (reflected in the slopes of the lines). In particular, the slopes of the lines in Figure 2 for the time pressure condition are especially flat, a tendency which corresponds with the very poor results for recognizing the navigation task when subjects are under time pressure. Essentially, since the speakers are trying to strip their utterances down to a bare minimum anyway, there is not much complex language processing that could be affected by a secondary task. Even the recognition accuracy when there is no time pressure is rather modest: After about 5 observations, the system assigns a probability of .60 to .65 to the correct hypothesis.

But note that it is not necessarily a problem if many users cope with a secondary task so well that it is difficult to recognize, on the basis of their speech, whether they are currently performing it or not. For these particular subjects, it may be less *important* to know whether they are performing a secondary task, since the secondary task may have little impact on their performance of other tasks (e.g., interacting with the mobile system). Further investigation of this issue will help to put the results just reported into perspective.

We could have made the navigation task easier to recognize simply by increasing its complexity in the experimental environment—perhaps to a point where it caused subjects' speech generation to break down completely. Instead, while developing the experiment we adjusted the level of complexity of the navigation task until it seemed typical of a situation in which a user is walking around a crowded airport while speaking into a device.

5 Summary of Contributions and Work in Progress

Our experiment differs from comparable previous experiments in (a) the number of independent variables examined simultaneously and (b) the relevance of the experimental tasks to mobile computing scenarios. A number of the specific effects identified had not been reported previously.

The evaluation of the learned user models is to our knowledge the first empirical evaluation of the feasibility of recognizing a person's time pressure and/or cognitive load on the basis of speech input.

We are currently pursuing the following extensions of this work:

- Use of more theoretically interpretable BN structures (cf. Wittig & Jameson, [11]) which will make it possible to analyze more clearly the reasons for particular aspects of the learned models' performance.
- Inclusion in the more articulate BN models of the remaining symptom variables that were recorded in the experiment but not included in the present study (cf. Section 3.1).

- Systematic studies of the reasons for the observed performance of the models (e.g., comparisons with tests in which some of the variables are omitted or not instantiated).

For practical use of these results, it will obviously be necessary to devise ways of coding the features of speech fully automatically, rather than largely manually as in the present study. Given that this goal is quite challenging for some of the features, our strategy has been to start by determining the diagnostic value of the features, so that the benefits of coding them automatically can be assessed. The results so far indicate that the features that would be most difficult to encode (e.g., content quality, self-corrections) have less diagnostic value than relatively easy features (e.g., duration of silent pauses, number of syllables).

References

1. André Berthold. Repräsentation und Verarbeitung sprachlicher Indikatoren für kognitive Ressourcenbeschränkungen [Representation and processing of linguistic indicators of cognitive resource limitations]. Master's thesis, Department of Computer Science, Saarland University, Germany, 1998.
2. André Berthold and Anthony Jameson. Interpreting symptoms of cognitive load in speech input. In Judy Kay, editor, *UM99, User Modeling: Proceedings of the Seventh International Conference*, pages 235–244. Springer Wien New York, Vienna, 1999.
3. Wray Buntine. A guide to the literature on learning probabilistic networks from data. *IEEE Transactions on Knowledge and Data Engineering*, 8:195–210, 1996.
4. Russell Greiner, Adam J. Grove, and Dale Schuurmans. Learning Bayesian nets that perform well. In Dan Geiger and Prakash P. Shenoy, editors, *Uncertainty in Artificial Intelligence: Proceedings of the Thirteenth Conference*, pages 198–207. Morgan Kaufmann, San Francisco, 1997.
5. Anthony Jameson, Barbara Großmann-Hutter, Leonie March, Ralf Rummer, Thorsten Bohnenberger, and Frank Wittig. When actions have consequences: Empirically based decision making for intelligent user interfaces. *Knowledge-Based Systems*, 14:75–92, 2001.
6. Anthony Jameson and Frank Wittig. Leveraging data about users in general in the learning of individual user models. In Bernhard Nebel, editor, *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*. Morgan Kaufmann, San Francisco, CA, 2001.
7. Sharon Oviatt. Multimodal interactive maps: Designing for human performance. *Human-Computer Interaction*, 12:93–129, 1997.
8. Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA, 1988.
9. Stuart J. Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice-Hall, Englewood Cliffs, NJ, 1995.
10. Ralph Schäfer and Thomas Weyrath. Assessing temporally variable user properties with dynamic Bayesian networks. In Anthony Jameson, Cécile Paris, and Carlo Tasso, editors, *User Modeling: Proceedings of the Sixth International Conference, UM97*, pages 377–388. Springer Wien New York, Vienna, 1997.
11. Frank Wittig and Anthony Jameson. Exploiting qualitative knowledge in the learning of conditional probabilities of Bayesian networks. In Craig Boutilier and Moisés Goldszmidt, editors, *Uncertainty in Artificial Intelligence: Proceedings of the Sixteenth Conference*, pages 644–652. Morgan Kaufmann, San Francisco, 2000.