Assessment of a User's Time Pressure and Cognitive Load on the Basis of Features of Speech

Anthony Jameson[1,4], Juergen Kiefer[2], Christian Müller[3], Barbara Großmann-Hutter[3], Frank Wittig[3], and Ralf Rummer[2⋆]

[1] German Research Center for Artificial Intelligence (DFKI), Saarbrücken, Germany
[2] Department of Psychology, Saarland University, Saarbrücken, Germany
[3] Department of Computer Science, Saarland University, Saarbrücken, Germany
[4] Fondazione Bruno Kessler – Istituto per Ricerca Scientifica e Tecnologica (FBK-irst), Trento, Italy

Abstract. One of the central questions addressed in the project READY was that of how a system can automatically recognize situationally determined resource limitations of its user—in particular, time pressure and cognitive load. This chapter summarizes most of the work done in READY on this topic, presenting as well some previously unpublished results. We first consider why on-line recognition or resource limitations can be useful by discussing the ways in which a system might adapt its behavior to perceived resource limitations. We then summarize a number of approaches to the recognition problem that have been taken in READY and other projects, before focusing on one particular approach: the analysis of features of a user's speech. In each of two similarly structured experiments, we created four experimental conditions that varied in terms of whether the user was (a) required to produce spoken utterances quickly or not; and (b) navigating within a simulated airport terminal or standing still. In the second experiment, additional distraction was caused by continuous loudspeaker announcements. The speech produced by the experimental subjects (32 in each experiment) was coded in terms of 7 variables. We report on the extent to which each of these variables was influenced by the subjects' resource limitations. We also trained dynamic Bayesian networks on the resulting data in order to see how well the information in the users' speech could serve as evidence as to which condition the user had been in. The results yield information about the accuracy that can be attained in this way and about the diagnostic value of some specific features of speech.

1 Introduction

The project READY (1996–2004) approached the topic of resource-adaptive cognitive processes from a different angle than most of the other projects represented in this

⋆ 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). Preparation of this manuscript was supported by the Province of Trento in its targeted research unit Prevolution (code PsychMM). The research benefited greatly from preparatory studies by André Berthold ([1]) and from advice by Werner Tack. Some results concerning Experiment 1 were described in a conference paper by Müller et al. ([2]).

volume: The resources in question were the cognitive resources of computer users; the adaptation was done by the system that they were using.

The type of system focused on in the research was mobile conversational systems, for reasons that will become clear below. The resource limitations of interest concerned the user's available time and working memory.

Since it would be impractical to discuss all of the lines of research in the project within a single chapter, this chapter will focus on one issue that was addressed in a number of studies over a period of several years, including one study whose results have not been published previously: the issue of how a system can estimate the time pressure and cognitive load of its user, in particular on the basis of evidence in the user's behavior with the system, such as their speech.

In passing, we will also mention some of the related work in the READY project, as well as other related research. Other aspects of the research in READY, especially concerning the use of probabilistic methods for user modeling, are discussed in the chapter by Wittig in this volume.

1.1 Reasons for Variation in Cognitive Load and Time Pressure

One salient issue in the design of mobile conversational interfaces is the role of situationally determined *resource limitations* of the user—specifically, time pressure and cognitive load.

Compared with the users of stationary interactive systems, mobile users are more likely to be experiencing environmentally induced cognitive load. The user $\mathcal{U}$'s attention to the environment may be due simply to distracting stimuli in the environment (as when $\mathcal{U}$ is being driven in a taxicab while using the system $\mathcal{S}$);[5] but often $\mathcal{U}$ will be attending actively to the environment while performing actions in it (e.g., handling objects or navigating through the environment). The tendency of users to attend to their environment and to multitask may be even greater with conversational mobile systems than with those that do not use speech as a communication channel, because of the largely eyes-free and hands-free character of speech.

Although users of stationary systems can of course also experience time pressure, especially acute time pressure can arise when a conversational interface is used during interaction with other persons or the environment. For example, a driver may want to complete a task while waiting at a stoplight; or a user may be interacting with another person who herself has little time available.

Research on how designers of technical devices can take situationally determined resource limitations into account has a long tradition in the field of engineering psychology (see, e.g., [3]). In the airplane cockpit, the automobile, or the nuclear power plant, the importance of factors like mental load and time pressure is too obvious to be overlooked. The research of this sort that seems most directly relevant to mobile conversational systems is research on in-car systems for drivers (see, e.g., [4]; [5]). The advent of conversational systems for drivers has been motivated largely by the perceived fundamental compatibility of speech with the task of driving (see, e.g., [6]).

[5] To simplify exposition, we will use the symbols $\mathcal{S}$ and $\mathcal{U}$ to denote a system and its user, respectively.

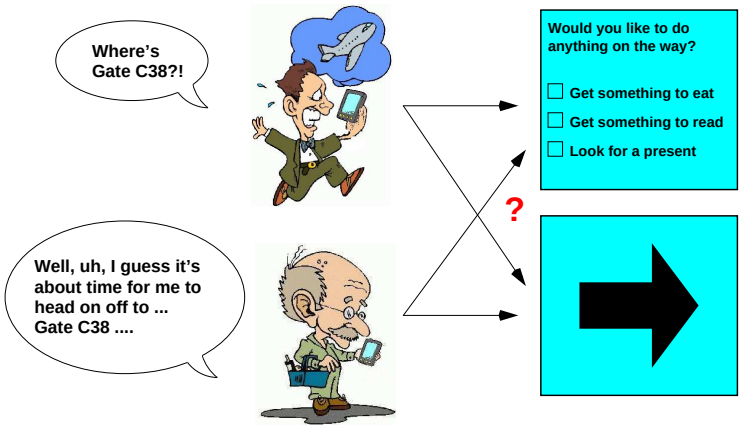

Fig. 1. Example of how a user's current resource limitations can call for different system responses. (Each of the two screens shown is a possible system response to the user's input utterance.)

With other types of mobile conversational interface, research on the role of user resource limitations is still in a relatively early stage. But it would be inappropriate to neglect them. Consider, for concreteness, the example of a conversational system that serves as an assistant to a traveler in a large airport, answering questions and providing guidance. Figure 1 illustrates how quite different system behaviors may be appropriate given different user resource limitations.

1.2 Why Automatic Adaptation?

There are, of course, straightforward ways of ensuring that a system shows appropriate behaviors in cases like this. First, the user could be allowed to specify explicitly what type of system response they prefer—for example, by including in the spoken query the request for a response that contains only the minimally necessary information. But especially when the user's resources are limited, such explicit specification may require too much mental effort and/or time. Second, the designers of the system can try to ensure that its basic design makes it highly usable even given severe resource limitations—for example, by providing only simple displays such as the lower one in Figure 1. But a design that is well suited for one particular combination of resource limitations may not be well suited to a different combination, or to a situation in which there are no significant limitations. For example, the minimalistic output on the lower screen in Figure 1 is unlikely to be optimal for the second user. And even the user experiencing time pressure might prefer a different type of display if he is not also experiencing cognitive load.

One possible approach to this dilemma is to give the system some capability to recognize the user's resource limitations automatically and to adapt to them with some degree of autonomy. In the next section, we will give some further examples of how this type of adaptation can be appropriate. Section 3 will then consider the first ques-

tion that this approach raises—How can a system automatically recognize resource limitations?—giving an overview of possible methods. Against this general background, the remaining major sections of the paper will present specific empirical results and analyses concerning the role of the user's speech as a source of evidence on which adaptation to resource limitations can be based.

2 Possible Forms of Adaptation

Let us suppose in this section that a mobile conversational interface $\mathcal{S}$ is capable of making some reasonably accurate estimate of the user $\mathcal{U}$'s resource limitations at a given moment. How might $\mathcal{S}$ make use of this assessment to generate more appropriate system behavior? If there are no plausible answers to this question, there is little point in investigating techniques for assessing resource limitations.

2.1 Interruption of Communication

Perhaps the simplest form of adaptation is for $\mathcal{S}$ simply to stop communicating with $\mathcal{U}$ when $\mathcal{S}$ perceives resource limitations. For example, [5] describes a prototype conversational in-car navigation system that interrupts its speech output whenever the driver applies the brakes. The goal is that in critical traffic situations, $\mathcal{U}$ should be able to devote their full attention to the driving task. In effect, the depression of the brake pedal is being interpreted as an indicator of high cognitive load.

2.2 Timing and Form of Notifications

Some conversational systems spontaneously present notifications to users. For example, the wearable NOMADIC RADIO ([7]) transmits audio messages (such as voice mail) to the user in a context-sensitive fashion. Although NOMADIC RADIO does not explicitly model $\mathcal{U}$'s cognitive load or time pressure, it does take into account related factors, such as whether $\mathcal{U}$ is currently interacting with $\mathcal{S}$ and whether $\mathcal{U}$ is in a meeting. In addition to postponing notifications, the system can choose from several forms of notification that have different degrees of obtrusiveness.

Other notification systems that assess the user's context have been presented by Horvitz and colleagues (see, e.g., [8]; [9]). These systems make use of decision-theoretic methods to weigh the benefits of a notification against the costs (e.g., distraction). Here again, cognitive load and time pressure are not modeled explicitly.

2.3 Dialog Strategy

Many conversational systems are capable of switching between different dialog styles depending on the current state of the interaction. For example, [10] describes TOOT, a prototype spoken dialog system for retrieving online train schedules. TOOT sometimes applies a highly conservative dialog strategy in which each piece of required information (e.g., destination, place of departure, time of departure) is elicited from the user through a focused question and then confirmed through a yes-no question. With less

conservative strategies, $\mathcal{S}$ asks more open questions that allow $\mathcal{U}$ to specify two or more pieces of information at a time (e.g., "How may I help you?"). $\mathcal{S}$ decides which strategy to use on the basis of features of the current dialog, such as the system's confidence in the success of its own speech recognition. The main motivation here is to allow users whose speech can be recognized relatively well to proceed through the dialog quickly, while still accommodating users whose speech is problematic. But analogous changes in dialog strategy could be based on assessments of cognitive load and/or time pressure: The more conservative strategies may be especially appropriate for users who are currently distracted by the environment or by another task, whereas they may be especially frustrating for users under time pressure.

Such hypotheses about the suitability of particular dialog styles for particular configurations of resource limitations of course require a theoretical and empirical foundation. An effort along these lines was made in a different line of research in the READY project ([11]): In an experimental setting, each of 24 subjects used a mouse to carry out spoken instructions regarding a graphical control panel (e.g., "Set *X* to 3, set *M* to 1, set *V* to 4"). In half of the trials, the instructions for a given panel were *bundled*, as in the example just given; in the other half of the trials, they were presented *stepwise*: After each single instruction (e.g., "Set *X* to 3"), the system waited until the user had completed the instruction and clicked on a confirmation button; then the system presented the next individual instruction. An orthogonal manipulation induced cognitive load in half of the trials through a secondary task that required subjects to attend to color changes in one part of the screen.

When instructions were presented bundled, subjects often made errors when a sequence comprised 3 or 4 instructions and when they were distracted by a secondary task. By contrast, the stepwise presentation of instructions was shown to be a slow but safe strategy, like the conservative dialog strategies discussed above: Subjects made very few errors even in the most difficult conditions. Given the assumption that users attach some value to both rapid task completion and the avoidance of errors, it can be shown that stepwise presentation is on the whole relatively suitable when $\mathcal{U}$ is experiencing cognitive load; but that the system's choice between the two modes should also be based on the length of the instruction sequence and the relative importance of execution speed and error avoidance. Although it was conducted in an artificial environment, this study empirically confirms the intuition that different dialog strategies can be suitable under different configurations of resource limitations.

2.4 Other Forms of Adaptation to Resource Limitations

Several other ways in which a conversational interface might adapt to resource limitations should be mentioned briefly for completeness, although they so far have been instantiated less clearly than the possibilities discussed above.

On the basis of perceived high cognitive load, a system might change its behavior as follows:

- Present a smaller amount of optional information that is not strictly required for the performance of $\mathcal{U}$'s system-related task.

For example, the airport assistant introduced above might stick to basic navigation instructions while guiding $\mathcal{U}$ from one location to another, leaving out information about airport facilities passed along the way.

– Present information in a style that is optimized for easy understanding, at the expense of other criteria (such as elegance or conciseness).

  Some stylistic features (e.g., simplicity and explicitness) are commonly recommended for texts that are typically read or heard by users who cannot be expected to be paying full attention, such as error messages and help texts (see, e.g., [3], chap. 6). The novel idea in an adaptive system is that the degree to which such elements should be included should depend on the perceived level of cognitive load, because of the tradeoffs with other criteria.

– Adapt the interface in such a way as to prevent errors that are typical of high cognitive load.

  A number of categories of *expert slip* are discussed by Norman ([12]), along with design remedies. Each such remedy (e.g., making objects more visually distinctive; asking for confirmation) tends to have some drawbacks. Since expert slips are especially likely when $\mathcal{U}$ is environmentally distracted, some remedies may become worthwhile under high cognitive load even if their drawbacks outweigh their advantages given low cognitive load.

Analogous suitable responses to time pressure might include the following:

– Present concrete instructions that describe specific actions, as opposed to encouraging $\mathcal{U}$ to discover procedures on her own or to form a robust mental model of the system.

– Optimize messages for speed of presentation and/or comprehension, if necessary at the expense of other criteria.

  For example, synthesized speech could be played at a faster rate, even though it might sound less pleasant and require more effort to understand.

3 Ways of Recognizing Resource Limitations

Given that there appears to be some potential benefit to the automatic recognition of a user's resource limitations, on the basis of what evidence can a system achieve such recognition?

3.1 Recognizing Likely Causes of Resource Limitations

A system may be able to recognize factors that tend to give rise to resource limitations in users. Any evidence that suggests the presence of such a factor constitutes indirect evidence for the corresponding resource limitation. Table 1 gives some examples of the many possibilities.

Table 1. Examples of ways in which an adaptive system might obtain information about causes of a user's resource limitations.

Cause of the resource limitation	Evidence of the cause that may be accessible to the adaptive system
Cognitive load	
Difficult driving situation	Information from navigation system
Use of a cognitively demanding interactive application	Information about applications currently being used by $\mathcal{U}$
Distracting noise and/or events in the environment	Sensing of the environment through microphones or cameras
Time pressure	
Requirement for fast task completion imposed by the environment (e.g., flight for which boarding is about to close)	$\mathcal{S}$'s access to information about environment−imposed constraints (e.g., boarding schedules)
Requirement for fast response imposed by $\mathcal{S}$ itself (e.g., instruction by $\mathcal{S}$ to perform a given action quickly)	$\mathcal{S}$'s access to its own processing history

3.2 Physiological Indicators

Within engineering psychology, there is a long tradition of research on physiological measures of cognitive load (see, e.g., [13]; [14]). Such measures have mostly been applied in laboratory or field studies, but there is some potential for using them for on-line recognition of and adaptation to cognitive load. Two relatively promising measures can serve as examples:

Heart Rate Variability Heart rate variability (see, e.g., [15]) tends to decrease with increasing overall mental workload. In a study somewhat similar in spirit to the one to be described in Sections 6, Rowe et al. ([15]) investigated the potential of heart rate variability to serve as in index of cognitive load, not only for the purpose of studying the workload induced by a given system but also for the purpose of allowing automatic adaptation. While this study did not yet yield clear conclusions about the value of heart rate variability for supporting on-line adaptation, they did suggest that further investigation of this possibility is warranted. Because of the need to attach electrodes to the user's body, heart rate variability does not fit especially naturally into the scenarios of mobile conversational interfaces; but perhaps ultimately the necessary sensors can be worn in an unobtrusive way and transmit data to a mobile device.

Pupil Diameter The diameter of a person's pupil has likewise been shown to vary systematically as a function of mental load—although it is also strongly affected by other factors, such as ambient illumination and the distance of objects being fixated (see, e.g., [16]). These other factors would be especially problematic with mobile systems. Pupil diameter can be measured with eye tracking equipment. With stationary system use, a

remote eye tracker can be used that does not have to be attached to the user's head—although the user is required to sit relatively still. For mobile use, a head-mounted eye tracker is required; for the time being, therefore, this type of measurement must be restricted to research studies, as opposed to normal system use. As is the case with heart rate variability, studies are required to determine whether and in what situations this type of information can play a useful role in a system that adapts to a user's resource limitations.

A study conducted within READY illustrated that that success is not guaranteed even in apparently optimal circumstances: In an experiment, Schultheis found no difference in the pupil diameter of subjects when they were reading very easy vs. very difficult texts on a computer screen (see, e.g., [17], [18]). A similar negative result was obtained by Iqbal et al. ([19]) on a simlar reading task, but these same authors obtained good accuracy results on different types of tasks.

Other Indices Other measures, which seem to have less immediate promise for use in mobile systems, include those that concern aspects of brain activity (for which, for example, Schultheis found some promising results in the experiment just mentioned; see also [20] for more recent and more promising results) and respiratory activity.

Comments One general advantage of physiological measures is that in general a continuous stream of data is received without the need for the user to produce any particular behavior solely for diagnostic purposes. Some measures, such as heart rate variability and pupil diameter, respond quickly enough to changes in cognitive load to make online adaptation in principle feasible. A general drawback is the need for specialized sensors, which users may find uncomfortable or restrictive.

3.3 Evidence in the User's Behavior With the System

A different general class of evidence comprises information about the user's behavior in interacting with the system—for example, $\mathcal{U}$'s use of manual input devices or $\mathcal{U}$'s speech. One positive aspect of these types of evidence is that special sensing devices may be unnecessary, because the information enters $\mathcal{S}$ through the normal input channels. Moreover, $\mathcal{U}$'s input behavior (e.g., the fact that $\mathcal{U}$ is making manual input errors or producing disfluent speech) may be of importance in its own right—that is, a fact that $\mathcal{S}$ might adapt to or take into account in its processing.

Evidence in the User's Motor Behavior Aspects of a user's motor behavior (e.g., tapping or dragging on a touchscreen with a stylus) could in principle reveal something about a user's resource limitations. A good deal of research has accumulated concerning features of motor behavior that typically arise under cognitive load and/or time pressure. Within the READY project, Lindmark ([21]) surveyed these relationships and suggested how they might be used for automatic recognition of resource limitations. For example, time pressure tends to lead to an increase in the stiffness of a person's limbs, which in turn tends to cause actions like tapping on the screen to be performed with relatively

high force ([22]); accordingly, when a given user employs more than the usual amount of force, this fact can be seen as suggestive evidence of time pressure. Cognitive load tends to increase the likelihood of expert slips (e.g., forgetting to perform an intended action; tapping on an icon that looks similar to the intended one; cf. [12]); if the system can recognize such an error as having been made—in general not a trivial task—it can use the error as evidence that suggests cognitive load. Some behaviors (such as the two just mentioned as examples) are made more likely by either cognitive load or time pressure. Therefore, any mechanism for interpreting such evidence will have to have some appropriate mechanism for adjusting its hypotheses concerning both of these resource limitations on the basis of the same evidence. Although the emphasis in the present chapter is not on inference mechanisms, one possible such mechanism will be discussed in connection with the analyses in Sections 7 and 8.

Evidence in the User's Speech With conversational interfaces, an especially natural type of indicator of resource limitations comprises features of the user's speech. Because $\mathcal{S}$ needs to process $\mathcal{U}$'s speech anyway, there must already exist some type of microphone for sensing the speech and some software for analyzing it. Therefore, as with motor indicators, in the best case the only further requirements concern software for identifying and interpreting the indicators. The prospects for recognizing resource limitations on the basis of this type of indicator will be examined in detail starting in Section 4.

4 Experiments: Introduction

As was argued in 3.3, features of a user's speech appear in several respects to be a promising source of information about a user's cognitive resource limitations. But an obvious first question is: Is there enough information available in a user's speech to support a reasonably reliable recognition of these resource limitations?

4.1 Earlier Research on Speech Indicators

Before initiating a time-consuming experimental study, we surveyed previously conducted studies of relations between cognitive load or time pressure and features of speech.[6]

Distinction From Other Topics The idea of making inferences about a speaker on the basis of features of their speech is by no means new. One topic of high practical importance is the recognition of emotion on the basis of speech (see, e.g., [34]). Part of this literature focuses on the effects of stress (see, e.g., [35]). Stress is related to cognitive load and time pressure, in that these resource limitations can be both causes and consequences of stress. But there are also essential aspects of the concept of *stress* that are not necessarily associated with cognitive load or time pressure: physiological

[6] Since this survey was made in 1998, it covered work through the late 1990s.

Table 2. Overview of the most important indicators of cognitive load found in some early studies.

Indicator	Direction*	Tally**	Example Study
Output rate			
Articulation rate	−	7/7	Lazarus−Mainka and Arnold (1987)
Speech rate	−	7/7	Kowal and O'Connell (1987)
Pauses			
Onset latency (duration)	+/(−)	9/11	Greene (1984)
Silent pauses (number)	+	4/5	Rummer (1996), Exp. 1 and 2
Silent pauses (duration, all)	+	6/8	Goldman−Eisler (1968)
Silent pauses (duration, intraphrasal only)	+	2/2	Butterworth (1980)
Filled pauses (number)	+	4/6	Wiese (1983)
Filled pauses (duration)	+	1/2	Grosjean and Deschamps (1973)
Indicators involving output quality			
Repetitions (number)	+	5/6	Deese (1980), Exp. 2
Sentence fragments (number)	+	4/5	Rummer (1996), Exp. 2
False starts (number)	+	2/4	Roßnagel (1995)
Self−corrections (number)***	+, −, 0	2, 1, 4	Oviatt (1995)

* "+" means that the measure was generally found to increase under conditions of high cognitive load; "−" means the opposite.

** *"m/n*" means that of *n* relevant studies, *m* found the tendency indicated in the second column. (In most cases the tendency was statistically significant.)

*** Results concerning self−corrections show an inconsistent pattern.

arousal and stressors such as noise or high acceleration (cf. [3], chap. 12). We believe that it can be important to be able to adapt to cognitive load or time pressure even when these factors are not present—for example, when the user is performing two tasks at once and would like to proceed quickly but is not especially concerned about the consequences of failure. We therefore focus here on previous studies that did not involve especially stressful situations. (A much more detailed and comprehensive analysis of studies like these is given by [1].)

Effects of Cognitive Load With regard to cognitive load, a number of features of speech have been investigated in multiple studies; hence it is possible to draw some fairly general conclusions concerning their dependence on cognitive load. Table 2 summarizes the most important of these indicators.

Effects of Time Pressure Perhaps surprisingly, the number of results that can be extracted from previous studies concerning the effects of time pressure is much smaller

than the number for cognitive load. One of the more obvious hypotheses is that people speak more quickly under time pressure. This hypothesis was confirmed in a study by Kelley and Stone ([36]), and a study by Marx ([37]) showed a marginal tendency of the same sort. This same study by Marx revealed a statistically significantly greater tendency of speakers who had been put under time pressure to repeat parts of utterances.

5 Experimental Method

5.1 Purpose of Experiments

The goals of our two experiments were (a) to fill the gap in knowledge concerning the impact of time pressure on features of speech; (b) to examine within a single setting a large number of features that had previously mostly been studied separately; and (c) to obtain raw data that could be used to determine how well cognitive load and time pressure can be recognized on the basis of speech.

We required some way of capturing users' speech while they are subject to known resource limitations. In principle it would be possible to capture the speech in fairly natural conditions, if we could confidently assess the resource limitations in these conditions. Healy and Picard ([38]) applied this strategy in their study of physiological assessment of driver stress: Subjects were required to drive along a route that included a number of events which had predictable stress levels.

We chose an experimental setting for our studies, so as to be able to exert greater control over both the independent variables and the nature of the speech utterances.

We conducted two experiments, separated by about 1 year in time; Experiment 2 can be seen as a replication and extension of Experiment 1. For concreteness, Experiment 1 will be described separately first.

5.2 Method for Experiment 1

Materials The experimental environment simulated a situation in which a user is walking through a crowded airport terminal while asking questions to a mobile assistance system via speech (see Figure 2). 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 ask a question, after motivating it with an introductory sentence. For example, for the picture shown in Figure 2, a subject might say "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.

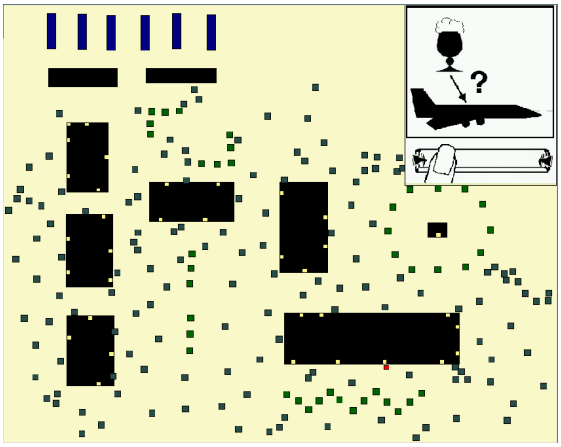

Fig. 2. Environment used in the experiments, with a typical pictorial stimulus.

The navigation task was designed to induce the sort cognitive load that would be induced by a nonverbal task performed by the user of a mobile system while interacting with the system. Walking around an airport would be one example of such a task; but there are of course differences between (a) walking in a real three-dimensional space and (b) moving an abstract figure within a two-dimensional computer screen. We do not refer to this condition as the "cognitive load" condition because it is not known to what extent the task actually induces cognitive load in any given subject.

– SPEECH 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.
  More specifically, in the condition with time pressure, the subject was told that his speech would be interpreted by an experienced airport assistant who was in great demand because of her extensive knowledge. Utterances directed to this assistant were to be completed quickly, so that she could go on to assist other airport visitors. In the condition without time pressure, subjects were to direct their utterances to a new, inexperienced airport assistant. In this condition, nothing was said about time limitations; the emphasis was to be on ensuring that this assistant understood the utterances.

The instructions concerning SPEECH TIME PRESSURE make it almost inevitable for some differences in the speech of the subjects to appear as a function of this variable. Still, there are empirical questions concerning (a) the particular forms that the utterances take in the two conditions (e.g., whether, under SPEECH TIME PRESSURE, subjects will articulate more quickly, use fewer words, and/or think less before starting to speak); and (b) whether the differences will be large enough to allow accurate discrimination between the two conditions.

We call this second variable SPEECH TIME PRESSURE to highlight its differences from other possible forms of time pressure. For example, if a person's goal is the quick completion of some larger task (e.g., getting to the departure gate), they may or may not try to save time by completing individual utterances quickly. But time pressure with regard to utterance completion can arise for various other reasons as well—for example, because of real or imagined time limitations on the part of the listener or system; because of a task that the user is performing that leaves only brief intervals free for speaking; or because of a high cost of utterances to the speaker, as in the case of an expensive communication channel. Any attempt to have a system adapt to SPEECH TIME PRESSURE in a given setting should take into account the likely reasons for this form of time pressure that might apply in that setting.

Procedure After an extensive introduction to the scenario, the environment, and the 4 (2×2) conditions, each subject dealt with 4 blocks of trials, each block involving 20 pictures 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 Each of the 2560 (32×80) utterances was transliterated and coded 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 sorted the stimulus pictures into 5 categories in terms of the complexity of the responses that they tended to call for. An aggregation of these ratings was later used to control for the different degrees of difficulty of the speech tasks invoked by the pictures.

In this chapter, we report results only for a subset of seven indicators which, on the basis of the results, seem most promising as indicators of cognitive load and/or time pressure:[7]

- NUMBER OF SYLLABLES: The number of syllables in the utterance.
- ARTICULATION RATE: The number of syllables articulated per second of speaking time, after elimination of the time for measurable silent pauses.
- 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). In accordance with usual practice, a silent pause is defined as a silence within the utterance that lasts for at least 200 ms.
- FILLED PAUSES: The corresponding measure for filled pauses (e.g., "Uhh").
- HESITATIONS: The number of silences with a duration of less than 200 ms, again relative to the length of the utterance in words.

[7] Much more detailed reports covering all of the variables are given by Müller ([39], for Experiment 1) and by Kiefer ([40], for Experiment 2).

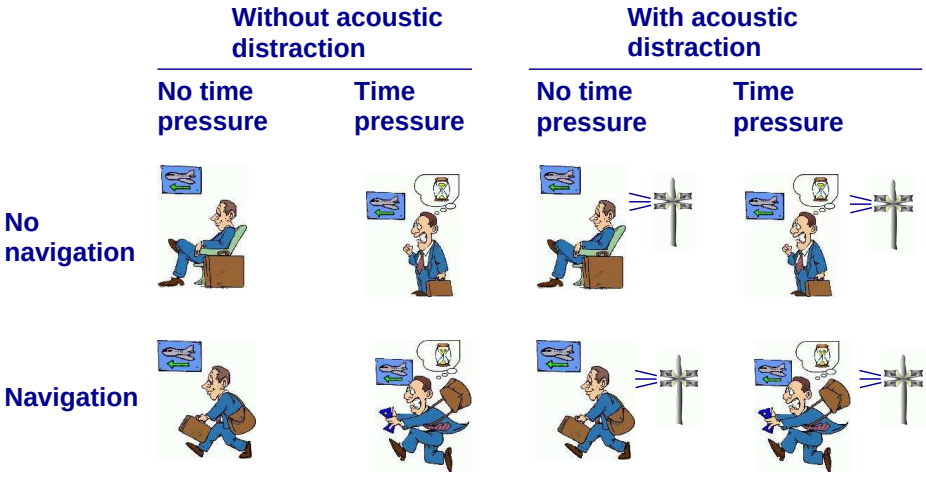

Fig. 3. Visualization of the eight conditions realized in Experiments 1 (left) and 2 (right). (In the experiments, the time pressure concerned specifically the time available to generate spoken input to the mobile system.)

- ONSET LATENCY: The length of the time interval between the presentation of the pictorial stimulus and the first syllable spoken by the subject.
- DISFLUENCIES: The logical disjunction of several binary variables, each of which indexes one type of speech disfluency: self-corrections involving either syntax or content; false starts; or interrupting speech in the middle of a sentence or a word. Although each of these variables has been treated as a separate dependent variable in some previous studies, they are grouped together here because each phenomenon in question occurs too infrequently in our data to give rise to statistically reliable effects. (Filled and silent pauses, which may also be regarded as disfluencies, are not counted here, because they are treated as separate variables.)

5.3 Method for Experiment 2

The method for Experiment 2 was identical to that for Experiment 1, with one exception: During all of the time in which a subject was performing the experimental tasks, they heard through a headphone prerecorded loudspeaker announcements of the sort that travelers typically hear at airport terminals (concerning matters such as flight departures, gate changes, missing persons, and security warnings). These German-language announcements, which had been recorded at Frankfurt Airport, were arranged digitally so that there were only minimal pauses between announcements. For our present purposes, the function of these announcements was to add an additional source of cognitive load—one which, in contrast to the navigation task, seemed likely to interfere more directly with the process of speech production, because of its verbal nature.

Figure 3 gives a graphical overview of the eight specific conditions that were realized in the two experiments. Our focus will be on the effects that occurred within each

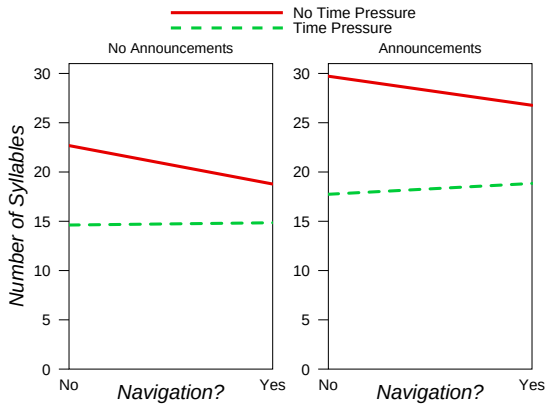

Fig. 4. Means for the variable NUMBER OF SYLLABLES in the four conditions of Experiment 1 (left) and Experiment 2 (right).

experiment. Although it is of some theoretical interest to see how the announcements affected speech production, in the present chapter we will not pay much attention to a comparison of the results with and without announcements. One reason is that there is little practical interest attached to the question of whether a system can recognize, on the basis of a user's speech, whether that user is being distracted by irrelevant speech from the environment: If $\mathcal{U}$'s speech can be picked up by a microphone, then presumably the presence of ambient speech could be directly detected via the microphone as well. Also, from a methodological point of view, we must be cautious in interpreting specific differences between the results of Experiments 1 and 2: Even though considerable effort was made to replicate the method of Experiment 1, for practical reasons Experiment 2 was conducted by a different experimenter and the utterances were transliterated by a different researcher. Moreover, the subjects were not necessarily sampled from the same population. It is therefore most realistic to focus on the robust results which are found in both of the experiments despite the differences between them.

6 Experimental Results

6.1 Statistical Analyses

For each of the indicators analyzed here, a three-way analysis of variance (ANOVA) was conducted, with two within-subject variables (NAVIGATION and SPEECH TIME PRESSURE) and one between-subject variable (ANNOUNCEMENTS).[8] In accordance

[8] Before the ANOVAs were conducted, multivariate analyses of variance had been conducted with a view to ensuring against capitalizing on chance with the relatively large number of ANOVAs; these MANOVAs demonstrated that the interpretation of the ANOVAs reported here is justified.

with the considerations just mentioned, we will interpret only the main effects of the within-subject variables and the interactions between them.

6.2 Number of Syllables

Figure 4 shows the means for the variable NUMBER OF SYLLABLES for each of the eight conditions. The ANOVA confirms that there is a highly significant main effect of SPEECH TIME PRESSURE ($F(1, 63) = 97.573, p < 0.001$): Not surprisingly, the instruction to finish each utterance quickly led to a much smaller number of syllables per utterance.

Somewhat less obviously, the requirement to navigate led to somewhat shorter utterances ($F(1, 63) = 8.295, p < 0.01$). Although there is no significant interaction between the two independent variables, the graphs suggest, plausibly, that the difference arises mainly in the condition without time pressure, in which the subjects were less ambitious with regard to the goal of producing unambiguous, high-quality utterances. When they were under time pressure, they were trying to keep their utterances short even when not navigating, so there was little room for the navigation task to cause further reduction in their length.

The results concerning NUMBER OF SYLLABLES are novel for the simple reason that previous studies have not in general included utterance length as a dependent variable. A likely reason for this omission is that utterance length has diagnostic significance only relative to a particular speech task: The fact that a user has produced a 15-syllable utterance in itself says little about her cognitive state; but if we know that the utterance was produced as an answer to a straightforward yes/no question, it may be significant. We will see in 7.1 how the properties of the current speech task can be taken into account in the interpretation of speech indicators.

6.3 Articulation Rate

As can be seen in Figure 5, on the average subjects produced more syllables per second when they were under time pressure than when they were not ($F(1, 63) = 47.726, p < 0.001$). Though this result is intuitively plausible, it is not logically necessary, given that there are other ways of coping with time pressure (cf. 4.1). There is also a tendency to articulate less quickly when navigating (see the slope of the two lines; $F(1, 63) = 4.355, p < 0.05$), as has been reported in a number of previous studies (cf. Table 2). This effect is stronger under time pressure; this interaction ($F(1, 63) = 5.565, p < 0.05$) is understandable in that, under time pressure, subjects are articulating relatively fast, so there is more room for them to slow down.

The fact that the two main effects and the interaction are statistically significant, even though the differences involving ARTICULATION RATE do not appear visually striking in the graphs, testifies to the precision and sensitivity of ARTICULATION RATE as an index.

6.4 Silent Pauses

The results for SILENT PAUSES (Figure 6) are complex. It is easily understandable that there is a highly significant main effect of SPEECH TIME PRESSURE ($F(1, 63) =$

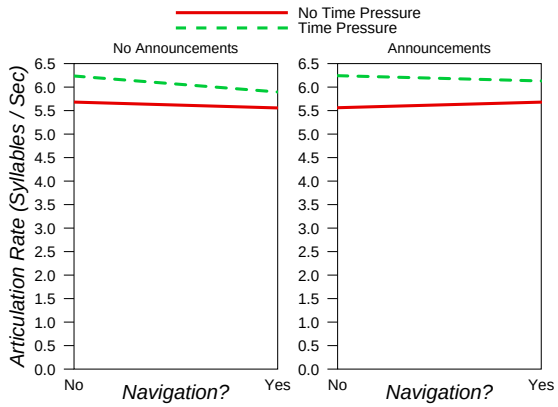

Fig. 5. Means for the variable ARTICULATION RATE in the four conditions of Experiment 1 (left) and Experiment 2 (right).

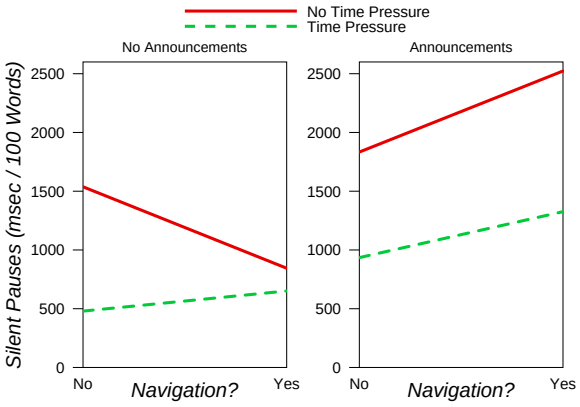

Fig. 6. Means for the variable SILENT PAUSES in the four conditions of Experiment 1 (left) and Experiment 2 (right).

$27.689, p < 0.001$): Without such pressure, subjects have no motivation to save time by avoiding pauses; perhaps even more importantly, they are motivated to produce high-quality utterances, which presumably tend to call for more careful planning, which can be accomplished during pauses. In particular, we have already seen (Figure 4) that utterances produced without time pressure tend to be considerably longer; and as was shown by Oviatt ([41]), longer utterances tend to be associated with a relatively high number of disfluencies such as silent pauses.

Regarding the effects of NAVIGATION, previous studies (cf. Table 2) had shown that a concurrent task tends to increase the number and/or length of silent pauses—plausibly enough, since a concurrent task demands the subjects' attention at least intermittently.

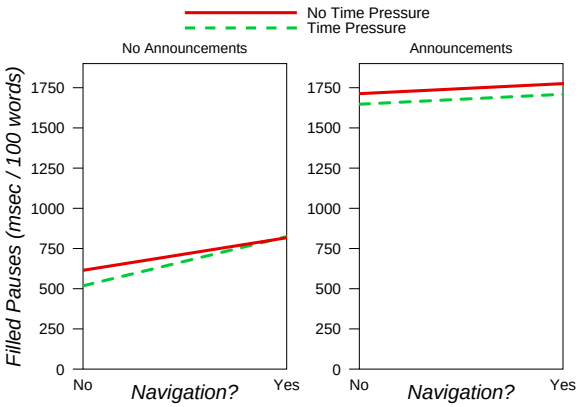

Fig. 7. Means for the variable FILLED PAUSES in the four conditions of Experiment 1 (left) and Experiment 2 (right).

This pattern is in fact seen in the upward slope of three of the four lines in Figure 6. The reason why there is no significant overall main effect of NAVIGATION is that a sharp decrease occurs in Experiment 1 when there is no time pressure. This decline is understandable when we recall that, without time pressure, the need to navigate leads to shorter utterances (Figure 4). In other words, subjects' adaptation to the navigation task proves more important in this case than the tendency of this task to increase cognitive load.

This specific result reminds us of a general point that is often emphasized in research on the effects of resource limitations on behavior (see, e.g., [3], chap. 11; [42]). Resource limitations do not in general have a direct and unavoidable impact on performance; typically, a person has some freedom to decide how to deal with them.

6.5 Filled Pauses

With the indicator FILLED PAUSES (Figure 7), the most striking difference between the two experiments appears. In Experiment 1 we see an effect that had been found in previous studies (cf. Table 2): an increase in filled pauses when a concurrent task is added. With the addition of the loudspeaker announcements in Experiment 2, this relatively subtle effect is reduced as the total duration of filled pauses increases by a factor of about 3; overall, there is no significant main effect of NAVIGATION. Although it is plausible that subjects generate more filled pauses in order to block out the distracting loudspeaker announcements, we should not attach much weight to this difference between the experiments, for the reasons given in 5.3.

6.6 Hesitations

The very short pauses counted by the variable HESITATIONS (Figure 8) occur significantly less frequently when the subject is navigating ($F(1, 63) = 8.407, p < 0.01$); a

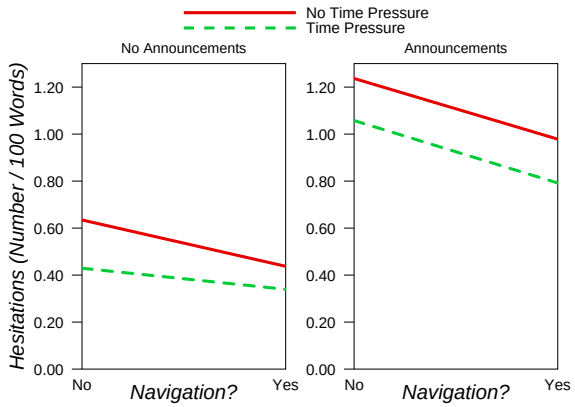

Fig. 8. Means for the variable HESITATIONS in the four conditions of Experiment 1 (left) and Experiment 2 (right).

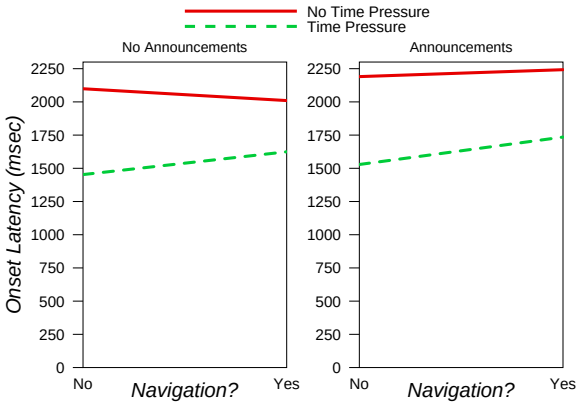

Fig. 9. Means for the variable ONSET LATENCY in the four conditions of Experiment 1 (left) and Experiment 2 (right).

possible explanation for this phenomenon is in terms of the reduction in the complexity of utterances when the subject is navigating (cf. Section 6.4). This result is novel in that virtually no previous studies have looked at hesitations as a dependent variable. The apparent effect of time pressure in the graphs is not statistically reliable, but note that it would be consistent with the results for SILENT PAUSES (6.4).

6.7 Onset Latency

Regarding ONSET LATENCY (Figure 9), we see a highly significant tendency for subjects to begin with the production of their utterance sooner when they have been in-

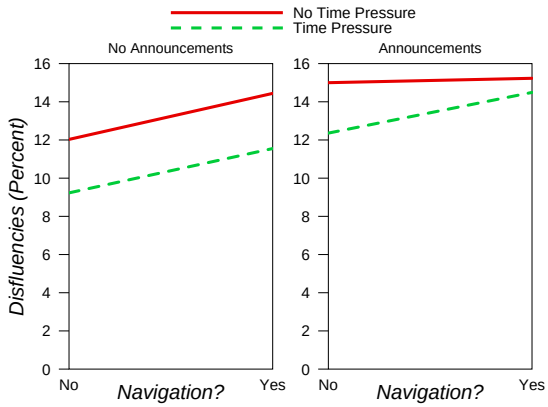

Fig. 10. Means for the variable DISFLUENCIES in the four conditions of Experiment 1 (left) and Experiment 2 (right).

structed to get finished with the utterance quickly ($F(1, 63) = 95.841, p < 0.001$). In addition to the obvious explanation that they are simply following instructions, this effect may be due in part to the lower complexity of the utterances produced under time pressure (cf. 6.2), which reduces the amount of planning required. The tendency (suggested by the lack of parallelism in the lines of each graph) for ONSET LATENCY to be affected more by NAVIGATION when there is SPEECH TIME PRESSURE is confirmed by a significant statistical interaction between the two independent variables ($F(1, 63) = 8.079, p < 0.05$). The positive impact of cognitive load on onset latency that was found in many previous studies (see Table 2) is not found here to a statistically significant degree, although there is a visible tendency in that direction.

6.8 Disfluencies

Although each of the specific types of disfluency summarized by the variable DISFLUENCIES occurs too infrequently to yield statistically significant differences as a function of the independent variables, a robust tendency does appear for the disjunction of the specific variables: As can be seen in Figure 10, DISFLUENCIES increase when the subject is required to navigate ($F(1, 63) = 8.403, p < 0.01$, as was shown in previous studies (cf. Table 2). The other tendency that is apparent in the figure—for disfluencies to increase when there is no time pressure—is not statistically reliable in these data, though it would be consistent with the greater complexity of utterances generated when there is no time pressure (cf. [41]).

6.9 Discussion

We have seen that, with the exception of FILLED PAUSES, each of the dependent variables discussed here shows one statistically reliable effect of time pressure and/or the

navigation task. As was mentioned above, some of these results replicate and extend findings from previous experimental research, while others yield new information—especially those that concern the independent variable of SPEECH TIME PRESSURE and its interactions with the presence of a concurrent task.

Taken together, these results suggest that observation of these variables in a person's speech might allow a system to infer that person's current resource limitations. But the question of the extent to which such recognition is possible is not directly addressed by the conventional analyses that we have presented so far: A statistically significant result in an ANOVA shows that the result is unlikely to have occurred because of chance alone, but it does not guarantee that the dependent variable in question will have diagnostic value. To determine the prospects for recognizing resource limitations, we will apply quite different methods in the following two sections.

7 Learning of User Models

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 main steps:

1. Use machine learning methods to create some sort of model relating resource limitations to speech indicators, using data such as those of these experiments (see the rest of this section).
2. Apply this model to the data of each user, using the features of their speech as evidence (Section 8).

7.1 Bayesian Network Structure

Regarding Step 1: There exists a great variety of machine learning techniques for classifying cases on the basis of their features, including support vector machines, neural networks, decision trees, and case-based reasoning.[9] A system that aims to recognize dynamically changing resource limitations imposes the following requirements on its learning and inference methods:

- The method should make it possible to interpret evidence from qualitatively different sources (cf. Section 3), ranging from likely causes of resource limitations to various types of indicator.
- The method should do justice to the fact that, while resource limitations change over time, the cognitive state of a user at any one moment will in most cases be similar to his or her state at the previous moment.
- The modeling method should yield a more or less interpretable model: Especially when several qualitatively different types of evidence are being used, it should be possible, by inspection of the model, to understand their relationships to one another (cf. [46]). Otherwise, it may be difficult to adapt the method to scenarios that involve different types of evidence.

[9] For general treatments of machine learning techniques, see [43]; [44]. Applications of such techniques to the modeling of computer users are discussed in [45].

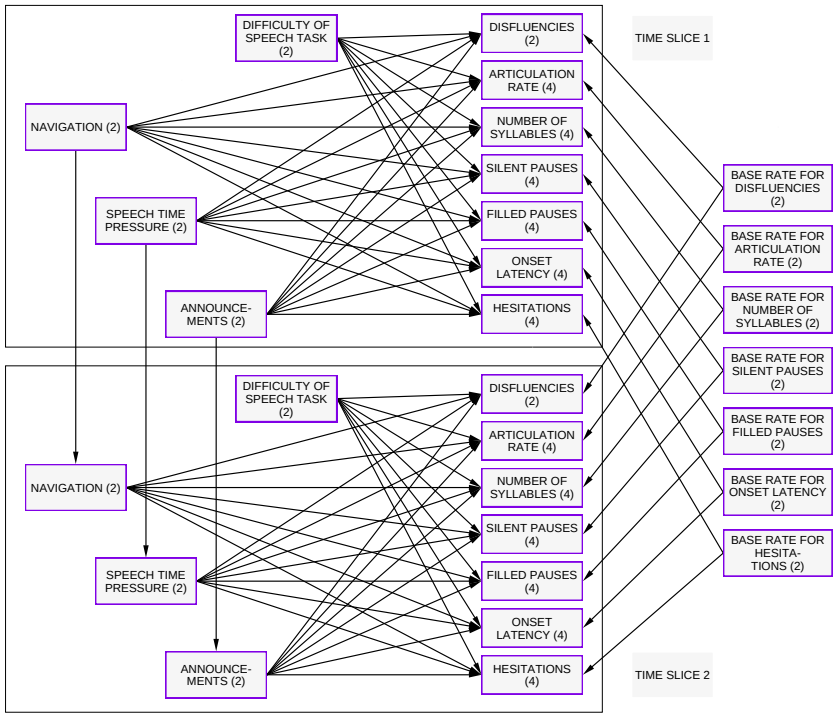

Fig. 11. Structure of the dynamic Bayesian network used in the evaluation of recognition accuracy. (Nodes within the two large boxes correspond to temporary variables that index features of the current utterance. Each number in parentheses shows the number of discrete states for the variable in question.)

– It should be possible to acquire a model of each individual user, so as to be able to take into account individual differences in the ways in which resource limitations are reflected in speech. But user model acquisition should also be able to take advantage of data acquired from users other than the current user, so that learning does not have to begin from scratch with each new user (cf. [47]).

Among the learning and inference techniques that best fit this combination of requirements are those that are associated with Bayesian networks (BNs).[10]

The BN structure employed in the present study is illustrated in Figure 11. (The nodes in the lower box labeled TIME SLICE 2 can be ignored for the moment.) We will first consider its qualitative structure; the quantitative modeling of the relationships among the variables represented will be discussed below.

The three nodes NAVIGATION, SPEECH TIME PRESSURE, and ANNOUNCEMENTS on the left correspond to the three main independent variables of the experiments. The

[10] The technical aspects of the use of Bayesian networks in the READY project, with a focus on the learning of BNs, are discussed in the chapter by Wittig in this volume.

node DIFFICULTY OF SPEECH TASK refers to the rated complexity of the speech task created by the stimulus picture (cf. Section 5). Each of these nodes represents a variable that can be seen as influencing the values of the seven dependent (indicator) variables that were analyzed in Section 6; these variables are represented by the seven nodes on the right within the box for TIME SLICE 1. Further influences on the indicator variables are represented by the seven nodes on the far right in the figure, which correspond to individual base rates for the seven indicator variables. They are introduced to take into account individual differences in the overall level of the indicator 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.

The BN structure in the figure shows a rather drastic simplification of the causal relationships that actually exist between the variables in question. For example, the absence of links among the base rate variables implies that these variables are statistically independent. In addition to being implausible, this assumption was shown to be false by our own factor analyses and applications of algorithms for learning BN structures from data. 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.[11]

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 what condition the user was in? Therefore, the variables NAVIGATION and SPEECH TIME PRESSURE can be viewed here as *static* variables whose value does not change over time. The seven 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*.[12]

7.2 Quantitative Parameters

In a BN such is the one used here, which does not include continuous variables, each variable has 2 or more discrete *states*, or possible values. For example, for NAVIGATION, the two states are "Navigation" and "No navigation". For the base rate variable BASE RATE FOR NUMBER OF SYLLABLES, each state corresponds to one of four ranges of numbers of syllables.

For each *root node* (i.e., a node that has no links directed at it), the system's initial expectation about the value of the variable in question is represented by a vector of probabilities that represents a probability distribution. For example, for each of the nodes SPEECH TIME PRESSURE, NAVIGATION, and ANNOUNCEMENTS, the probabilities are simply <.50, .50>, reflecting the fact that each value of each of these variables

[11] 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.

[12] An explanation of the general principles of dynamic Bayesian networks can be found, for example, in chap. 17 of [48]. A discussion with regard to user modeling of the sort done here is given by [49].

occurred equally often in the experiments. For each of the base rate nodes, the probability vector reflects the empirically determined distribution of the base rate in question in the group of subjects in these experiments.

For each node that is not a root node, a *conditional probability table* (CPT) represents the system's assumptions about how the value of the variable is related to the values of its *parent variables* (corresponding to the nodes with links that point to it). For example, each probability in the CPT for DISFLUENCIES represents the likelihood that a disfluency will occur (or not occur) in an utterance, given particular values of the parent variables SPEECH TIME PRESSURE, NAVIGATION, ANNOUNCEMENTS, DIFFICULTY OF SPEECH TASK, and BASE RATE FOR DISFLUENCIES.

A BN makes probabilistic inferences when it is evaluated: Typically, one or more variables in the BN are *instantiated*; that is, the probability distribution representing the system's belief about the value of such a variable is replaced by a probability distribution which expresses certainty that one particular value is realized. Then the BN is reevaluated; typically the system's beliefs about some of the uninstantiated variables are updated to be consistent with the new information provided by the instantiations.

7.3 Learning the Quantitative Parameters

Although we specified the structure of the BN shown in Figure 11 by hand, the probabilities need to be learned empirically. Such learning is quite straightforward in a BN (such as this one) that includes only observable variables: In accordance with the usual maximum-likelihood method (see, e.g., [50]), the estimate of each (conditional) probability is computed simply in terms of the (relative) frequencies in the data.[13]

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 for a separate BN using the data from the other 63 subjects. The learned BN has the structure shown in Figure 11 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.

8 Evaluation of the User Models

8.1 Procedure

The basic idea of the evaluation of the learned models can be explained with reference to Figure 3: Given the behavior of a subject in one of the eight experimental conditions, our system will try to infer which condition the subject was in when he or she produced that behavior. More specifically, when asking the system to assess the probability that $\mathcal{U}$ was under time pressure, we will tell the system whether $\mathcal{U}$ was navigating and whether $\mathcal{U}$ was distracted by loudspeaker announcements. Similarly, when asking the system to assess the probability that $\mathcal{U}$ was navigating, we will specify the true values of the other two independent variables. (We will not report on tests of how well $\mathcal{S}$ can discriminate

[13] The learning of BNs in much more complex settings is discused in the chapter by Wittig in this volume.

Table 3. Procedure used in evaluating the accuracy with which a learned Bayesian network assesses the value of the variable SPEECH TIME PRESSURE for a given user. (The procedure is identical when the value of NAVIGATION is to be assessed, except that the roles of *T* and *N* are interchanged.)

Relevant variables and their values

- A user $\mathcal{U}$
- Values *t*, *n*, and *a* of the Boolean variables *T* (Speech Time Pressure), *N* (Navigation), and *A* (Announcements)

Task

Infer the value of *T* on the basis of indicators in $\mathcal{U}$'s speech

Preparation of the test data

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

Evaluating recognition accuracy

Initialize the model:

1. Create the first time slice of the BN for $\mathcal{U}$
2. Instantiate each of the individual base rate variables with its true value for $\mathcal{U}$
3. Also instantiate *N* and *A* with their true values *n* and *a*, but leave the variable *T* (whose value is to be inferred) uninstantiated

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

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

between the presence and the absence of ANNOUNCEMENTS, for the reasons given in 5.3, except to note in passing that the results are roughly comparable to those reported below for the recognition of NAVIGATION.)

More formally, the procedure for evaluating a learned BN is given in Table 3.

8.2 Results

Because of the differences between Experiments 1 and 2 (cf. 5.3), in Figure 12 the results of the modeling evaluation are shown separately for each of the two experiments. Each curve is the result of averaging 32 curves, one for each subject in the experiment in question.[14]

Recognizing Time Pressure Looking first at the results for recognizing SPEECH TIME PRESSURE (left-hand graphs), we see that the BNs are on the whole rather successful: The average probability assigned to the actual current condition rises sharply during

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

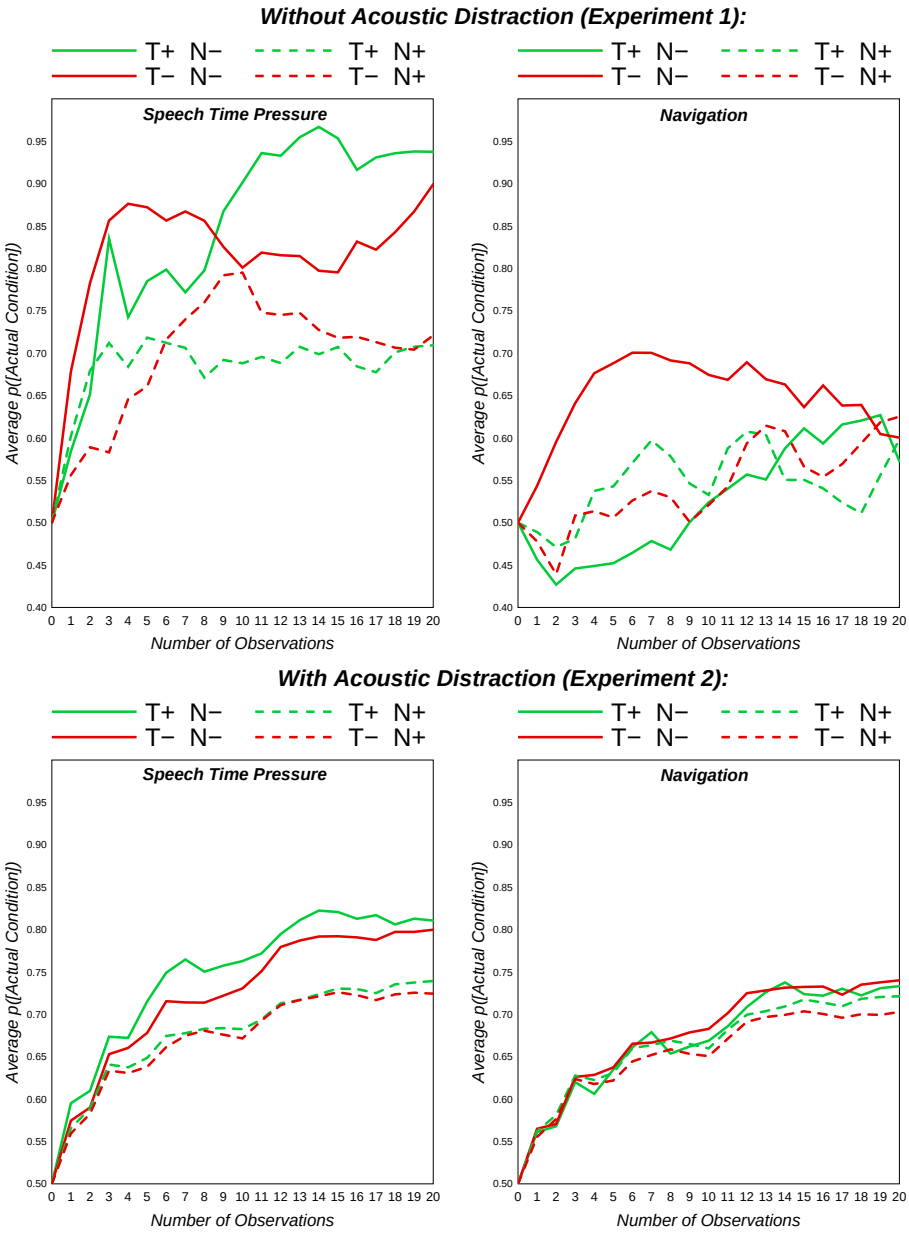

Fig. 12. Accuracy of the learned dynamic Bayesian networks in inferring the correct value of SPEECH TIME PRESSURE ("T", left) and NAVIGATION ("N", right) in Experiment 1 (above) and Experiment 2 (below). (Each curve shows the aggregated results for one combination of values of the variables SPEECH TIME PRESSURE, NAVIGATION, and ANNOUNCEMENTS. In each curve, the point for the *i*th observation shows the average probability which the Bayesian network assigned to the subject's actual condition after processing the first *i* observations.)

the first few observations. Note that in each experiment, recognition of SPEECH TIME PRESSURE is easier when there is no navigation task.[15] This result is understandable in the light of the conventional analyses discussed in Section 6: On the whole, the effects of time pressure were somewhat greater when there was no navigation task (i.e., the lines tended to be farther apart on the left-hand sides of the graphs), since in that condition speakers were able to respond more sensitively to the time pressure (or lack of it).

Recognizing Navigation In Experiment 2, the results for recognition of NAVIGATION are consistent over the four conditions: After several observations, the system on the average assigns a probability of roughly 0.65 to the correct condition. The fact that this probability never rises much above 0.70, even after 20 observations, shows that there is an inherent difficulty in discriminating between the presence and absence of NAVIGATION which cannot be overcome through the provision of a large number of observations.

In Experiment 1, the results are generally poorer, and they show rather strange variations between conditions and over time.[16] One way of understanding the better results for Experiment 2 is simply to note that the indicators shown in Figures 6 through 10 tend to occur to a greater extent in Experiment 2 (i.e., the lines in the right-hand graphs in these figures tend to be higher than those in the left-hand graphs). Since these indicators are on the whole low-frequency events, any increase in their frequency is likely to make recognition more accurate. It may be speculated that this overall difference in the frequency of indicators is due to the presence of loudspeaker announcements in Experiment 2, which push subjects closer to the limits of their processing capacity.

Dispensing With Individual Indicators Especially when we consider the practical problem of measuring indicators automatically (see 8.3), it becomes interesting to consider which of the seven indicators might be dispensable on the grounds that they do not add significantly to the accuracy of recognition. We repeated the simulations summarized in Figure 12 seven times, each time leaving out one of the seven indicators. Since it would be tedious and imprecise to examine seven further sets of four graphs similar to those shown in Figure 12, we computed for each graph a single number that summarizes the success of recognition: the mean of the 80 probabilities shown in the four curves of the graph. The question then becomes: To what extent do these mean probabilities decline when one of the indicator variables is left out of consideration?

Table 4 shows the results. The indicator whose removal has the greatest impact is clearly NUMBER OF SYLLABLES. Each of the other indicators seems surprisingly

[15] Since this statement applies to each of the observations 1 through 20 in each experiment, the difference is statistically reliable for each experiment with $p < .001$ by a sign test.

[16] As was mentioned in an earlier report on Experiment 1 ([2]), the results for the recognition of navigation are actually better if the system is *not* told whether $\mathcal{U}$ was under time pressure—perhaps because the BN then bases its assessment on a larger number of conditional probabilities and hence, indirectly, on a larger amount of data from other subjects. Overall, however, there is no systematic tendency for recognition to be better or worse when the system is told the value of the independent variable(s) that it is not trying to assess.

Table 4. Impact on recognition accuracy of leaving out of consideration each of the seven indicator variables. (Each number in the column "All Indicators" is the mean of the 80 probabilities shown in the corresponding graph in Figure 12, expressed as a percentage. Each number in a column to the right (except the rightmost column) shows the corresponding mean change in accuracy (as a percentage, but in absolute terms) when the simulation is performed without use of the indicator variable in question. The rightmost column shows the sum of these changes.)

	All Indicators	Syllables	Articulation Rate	Filled Pauses	Hesitations	Onset Latency	Disfluencies	Silent Pauses	Sum of Changes
Speech Time Pressure:									
Experiment 1	75.76	−6.13	.00	−.09	+.07	−3.61	+.36	+1.12	−8.29
Experiment 2	70.32	−5.86	−1.17	−1.21	−.60	−.18	−.04	−.04	−9.10
Navigation:									
Experiment 1	56.58	−1.86	−.66	−2.02	−.29	+1.39	+.12	−.99	−4.31
Experiment 2	66.51	−4.94	−1.11	−.84	−.54	−.35	−.03	+.44	−7.39

dispensable; and in a few cases leaving an indicator out even improves recognition accuracy. As the final column shows, the sum of the changes that result from leaving individual indicators out is much smaller than the extent to which recognition exceeds the chance level of 50%. This fact shows that the contributions of the indicators are not simply additive: It may be possible to leave out one indicator without much loss of accuracy because the information that it contributes is largely supplied by other indicators; but it would not be advisable to leave out all or most of them.

The indicator that it would presumably be most practically useful to omit is DISFLUENCIES: Automatically recognizing linguistic phenomena such as self-corrections, false starts, and interrupted sentences is considerably more difficult than measuring (silent or filled) pauses and counting syllables, which is all that is required for the other indicators.[17] As Table 4 shows, the variable DISFLUENCIES adds at best negligible value, provided that the other indicators are available.

8.3 Discussion

One question concerns the extent to which the results concerning the recognition of resource limitations can be generalized to different (and more realistic) settings. Certainly the specific probabilities of correct recognition are dependent on features of the particular situation—witness the differences that arose even between these two very similar experiments. For our analyses, it was certainly helpful that the experimental situation was highly constrained. Moreover, it was important for the system to know the difficulty of the specific speech task that the user was performing. In an interactive system, the corresponding information would consist in expectations about the complexity of

[17] Portable hardware (with associated software) for detecting and analyzing pauses in speech is commercially available.

the utterance that the user is likely to produce in any given situation (for example, after a question about the user's desired destination).

In sum, much work remains to be done before features in a user's speech can be used for the recognition of the resource limitations of a real user of an interactive system; and even in the long run this possibility will probably be subject to various restrictions—for example, concerning the predictability of the speech produced by users.

9 Summary of Contributions and Remaining Work

One goal of the present chapter was to provide a framework for thinking about the prospects for adapting to a user's cognitive resource limitations in interactive systems in general and in mobile conversational systems in particular. We discussed why such adaptation might be worthwhile, what forms it might take, and how the resource limitations might be automatically assessed.

The more specific goal was to explore the prospects of exploiting the user's speech as a source of evidence for the recognition of resource limitations. One respect in which the two experiments presented differ from comparable previous experiments concerns the number of independent variables examined simultaneously: Whereas almost all previous studies had examined the effects of just one variable (usually cognitive load), our experiments orthogonally manipulated cognitive load and speech time pressure, as well as repeating the experiment with and without distraction from irrelevant speech. The nature of the manipulations makes the experiments somewhat more relevant to scenarios of mobile conversational interaction than previous experiments were. But the most important new contribution concerns the results on the diagnostic value of seven specific features of speech: The evaluation experiments show that these indicators together do permit a degree of recognition of time pressure and cognitive load that could be useful in some situations, and they indicate the effects of leaving out individual features that would be relatively hard to recognize automatically.

Any attempt to apply the ideas and results from this chapter in a particular application scenario will necessarily involve considerable further work and creativity. But we believe that the results presented here will be helpful as a starting point.

References

1. Berthold, A.: 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 (1998)
2. Müller, C., Großmann-Hutter, B., Jameson, A., Rummer, R., Wittig, F.: Recognizing time pressure and cognitive load on the basis of speech: An experimental study. In Bauer, M., Gmytrasiewicz, P., Vassileva, J., eds.: UM2001, User Modeling: Proceedings of the Eighth International Conference. Springer, Berlin (2001) 24–33
3. Wickens, C.D., Hollands, J.G.: Engineering Psychology and Human Performance. 3rd edn. Prentice Hall, Upper Saddle River, NJ (2000)
4. Wierwille, W.W.: Visual and manual demands of in-car controls and displays. In Peacock, B., Karwowski, W., eds.: Automotive Ergonomics. Taylor and Francis, London (1993) 299–320

5. Bernsen, N.O., Dybkjær, L.: Exploring natural interaction in the car. In: Proceedings of the CLASS Workshop on Natural Interactivity and Intelligent Interactive Information Representation, Verona, Italy (2001)
6. Salvucci, D.D.: Predicting the effects of in-car interfaces on driver behavior using a cognitive architecture. In Jacko, J.A., Sears, A., Beaudouin-Lafon, M., Jacob, R.J., eds.: Human Factors in Computing Systems: CHI 2001 Conference Proceedings. ACM, New York (2001) 120–127
7. Sawhney, N., Schmandt, C.: Nomadic Radio: Speech and audio interaction for contextual messaging in nomadic environments. ACM Transactions on Computer-Human Interaction **7** (2000) 353–383
8. Horvitz, E.: Principles of mixed-initiative user interfaces. In Williams, M.G., Altom, M.W., Ehrlich, K., Newman, W., eds.: Human Factors in Computing Systems: CHI 1999 Conference Proceedings. ACM, New York (1999) 159–166
9. Horvitz, E., Jacobs, A., Hovel, D.: Attention-sensitive alerting. In Laskey, K.B., Prade, H., eds.: Uncertainty in Artificial Intelligence: Proceedings of the Fifteenth Conference. Morgan Kaufmann, San Francisco (1999) 305–313
10. Litman, D.J., Pan, S.: Designing and evaluating an adaptive spoken dialogue system. User Modeling and User-Adapted Interaction **12**(2–3) (2002) 111–137
11. Jameson, A., Großmann-Hutter, B., March, L., Rummer, R., Bohnenberger, T., Wittig, F.: When actions have consequences: Empirically based decision making for intelligent user interfaces. Knowledge-Based Systems **14** (2001) 75–92
12. Norman, D.A.: Design rules based on analyses of human error. Communications of the ACM **26** (1983) 254–258
13. Kramer, A.F.: Physiological metrics of mental workload: A review of recent progress. In Damos, D.L., ed.: Multiple-Task Performance. Taylor and Francis, London (1991) 279–328
14. Wilson, G.F., Eggemeier, F.T.: Psychophysiological assessment of workload in multi-task environments. In Damos, D.L., ed.: Multiple-Task Performance. Taylor and Francis, London (1991) 329–360
15. Rowe, D.W., Silbert, J., Irwin, D.: Heart rate variability: Indicator of user state as an aid to human-computer interaction. In Karat, C.M., Lund, A., Coutaz, J., Karat, J., eds.: Human Factors in Computing Systems: CHI 1998 Conference Proceedings. ACM, New York (1998) 480–487
16. Granholm, E., Asarnow, R.F., Sarkin, A.J., Dykes, K.L.: Pupillary responses index cognitive resource limitations. Psychophysiology **33**(4) (1996) 457–461
17. Schultheis, H.: Pupillengröße und kognitive Belastung [Pupil size and cognitive load]. Master's thesis, Saarland University, Department of Psychology (2004)
18. Schultheis, H., Jameson, A.: Assessing cognitive load in adaptive hypermedia systems: Physiological and behavioral methods. In Nejdl, W., De Bra, P., eds.: Adaptive Hypermedia and Adaptive Web-Based Systems: Proceedings of AH 2004. Springer, Berlin (2004) 225–234
19. Iqbal, S.T., Zheng, X.S., Bailey, B.P.: Task-evoked pupillary response to mental workload in human-computer interaction. In: Extended Abstracts for CHI 2004, Vienna (2004) 1477–1480
20. Grimes, D., Tan, D.S., Hudson, S.E., Shenoy, P., Rao, R.P.: Feasibility and pragmatics of classifying working memory load with an electroencephalograph. In Burnett, M., Costabile, M.F., Catarci, T., de Ruyter, B., Tan, D., Czerwinski, M., Lund, A., eds.: Human Factors in Computing Systems: CHI 2008 Conference Proceedings. ACM, New York (2008) 835–844
21. Lindmark, K.: Interpreting symptoms of cognitive load and time pressure in manual input. Master's thesis, Department of Computer Science, Saarland University, Germany (2000)
22. van Galen, G.P., van Huygevoort, M.: Error, stress and the role of neuromotor noise in space oriented behaviour. Biological Psychology **51** (2000) 151–171

23. Lazarus-Mainka, G., Arnold, M.: Implizite Strategien bei Doppeltätigkeit: Sprechen = Zuhören und Sortieren [Implicit strategies in dual tasks: Speaking = listening and sorting]. Zeitschrift für experimentelle und angewandte Psychologie **34** (1987) 286–300
24. Kowal, S., O'Connell, D.: Some temporal aspects of stories told while or after watching a film. Bulletin of the Psychonomic Society **25** (1987) 364–366
25. Greene, J.O.: Speech preparation processes and verbal fluency. Human Communication Research **11** (1984) 61–84
26. Rummer, R.: Kognitive Beanspruchung beim Sprechen [Cognitive load in speaking]. Beltz, Weinheim, Germany (1996)
27. Goldman-Eisler, F.: Psycholinguistics: Experiments in Spontaneous Speech. Academic Press, London (1968)
28. Butterworth, B.: Evidence from pauses in speech. In Butterworth, B., ed.: Language Production. Academic Press, New York, London (1980) 155–176
29. Wiese, R.: Psycholinguistische Aspekte der Sprachproduktion [Psycholinguistic Aspects of Speech Production]. PhD thesis, Hamburg (1983)
30. Grosjean, F., Deschamps, A.: Analyse des variables temporelles du français spontané [analysis of temporal variables in spontaneous French]. Phonetica **28** (1973) 191
31. Deese, J.: Pauses, prosody, and the demands of production in language. In Dechert, H.W., Raupach, M., eds.: Temporal Variables in Speech: Studies in Honour of Frieda Goldman-Eisler. Mouton, The Hague (1980) 69–84
32. Roßnagel, C.: Übung und Hörerorientierung beim monologischen Instruieren: Zur Differenzierung einer Grundannahme [Practice and listener-orientation in the delivery of instruction monologs: Differentiation of a basic assumption]. Sprache & Kognition **14** (1995) 16–26
33. Oviatt, S.: Predicting spoken disfluencies during human-computer interaction. Computer Speech and Language **9** (1995) 19–35
34. Banse, R., Scherer, K.R.: Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology **70**(3) (1996) 614–636
35. Fernandez, R., Picard, R.W.: Modeling drivers' speech under stress. In: Proceedings of the ISCA Workshop on Speech and Emotions, Belfast (2000)
36. Kelly, K.R., Stone, G.L.: Effects of time limits on the interview behaviour of Type A and B persons within a brief counseling interview. Journal of Counseling Psychology **29** (1982) 454–459
37. Marx, E.: Über die Wirkung von Zeitdruck auf Sprachproduktionsprozesse [The Effect of Time Pressure on Speech Production Processes]. PhD thesis, University of Münster, Germany (1984)
38. Healey, J., Picard, R.: SmartCar: Detecting driver stress. In: Proceedings of the Fifteenth International Conference on Pattern Recognition, Barcelona (2000) 4218–4221
39. Müller, C.: Symptome von Zeitdruck und kognitiver Belastung in gesprochener Sprache: eine experimentelle Untersuchung [Symptoms of time pressure and cognitive load in speech: An experimental study]. Master's thesis, Department of Computational Linguistics, Saarland University, Germany (2001)
40. Kiefer, J.: Auswirkungen von Ablenkung durch gehörte Sprache und eigene Handlungen auf die Sprachproduktion [Effects on speech production of distraction through overheard speech and one's own actions]. Master's thesis, Department of Psychology, Saarland University, Germany (2002)
41. Oviatt, S.: Multimodal interactive maps: Designing for human performance. Human-Computer Interaction **12** (1997) 93–129
42. Baber, C., Mellor, B.: The effects of workload on speaking: Implications for the design of speech recognition systems. In: Contemporary Ergonomics: Proceedings of the Annual Conference of the Ergonomics Society. (1996) 513–517

43. Langley, P.: Elements of Machine Learning. Morgan Kaufmann, San Francisco (1996)
44. Mitchell, T.M.: Machine Learning. McGraw-Hill, Boston (1997)
45. Webb, G., Pazzani, M.J., Billsus, D.: Machine learning for user modeling. User Modeling and User-Adapted Interaction **11** (2001) 19–29
46. Wittig, F., Jameson, A.: Exploiting qualitative knowledge in the learning of conditional probabilities of Bayesian networks. In Boutilier, C., Goldszmidt, M., eds.: Uncertainty in Artificial Intelligence: Proceedings of the Sixteenth Conference. Morgan Kaufmann, San Francisco (2000) 644–652
47. Jameson, A., Wittig, F.: Leveraging data about users in general in the learning of individual user models. In Nebel, B., ed.: Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence. Morgan Kaufmann, San Francisco (2001) 1185–1192
48. Russell, S.J., Norvig, P.: Artificial Intelligence: A Modern Approach. 2nd edn. Prentice-Hall, Englewood Cliffs, NJ (2003)
49. Schäfer, R., Weyrath, T.: Assessing temporally variable user properties with dynamic Bayesian networks. In Jameson, A., Paris, C., Tasso, C., eds.: User Modeling: Proceedings of the Sixth International Conference, UM97. Springer Wien New York, Vienna (1997) 377–388
50. Buntine, W.: A guide to the literature on learning probabilistic networks from data. IEEE Transactions on Knowledge and Data Engineering **8** (1996) 195–210