Jaakko Hakulinen, Markku Turunen,
and Kari-Jouko Räihä

Tutoring in a Spoken Language Dialogue System

TIETOJENKÄSITTELYTIETEIDEN LAITOS
TAMPEREEN YLIOPISTO

A-2006-3

TAMPERE 2006

Tutoring in a Spoken Language Dialogue System

JAAKKO HAKULINEN, MARKKU TURUNEN, and KARI-JOUKO RÄIHÄ
Tampere Unit for Computer-Human Interaction
Department of Computer Sciences
University of Tampere, Finland

We have developed interactive software tutors to teach users how to use a spoken dialogue timetable system. The tutors teach the functionality and interaction style of the telephone-based timetable system to new users by guiding users and monitoring their interaction. The primary modality of the tutors is graphics and they feature a visual representation of the spoken dialogue between a user and the system. Two different versions of tutoring were compared to a static web manual with the same information in a between-subjects experiment with 27 participants. Participants' evaluations of guidance materials were the most positive towards a tutor featuring a graphical interface representation of the timetable query. An otherwise similar tutor, which did not have the graphical user interface representation, received the weakest evaluations. Error rate variances suggest that tutoring is better than static guidance especially for those who most need guidance.

Categories and Subject Descriptors: *User Interfaces*; H5.2 [**Information Interfaces and Presentation**]: User Interfaces - *Voice I/O, Natural language, Training, help and documentation*;
General Terms: Human Factors, Design, Documentation, Experimentation
Additional Key Words and Phrases: Human-computer interaction, software tutoring, spoken dialogue systems, speech user interfaces, user guidance, multimedia.

1. INTRODUCTION

Advances in speech technology have made it possible to implement useful speech-based applications. Growing emphasis on timely customer services and increased safety provided by in-vehicle information systems, for instance, have radically increased the user base of speech-based systems and services. The take-up of more flexible spoken dialogue systems has been slower, though they, too, have been in public use for more than two decades [Cox et al. 2000].

The challenges of designing spoken dialogue systems are well known, as are the usual solutions. How do the users know the functionality provided by a speech-based system? How do they know what to say to the system? How do they know when to speak? A well designed speech interface supports the users' natural way of speaking. However, in practice the interface must also guide users to speak in a way that the system is able to understand. Implicit and explicit prompts, hints, and tapering are commonly used methods for this. They embed the guidance in the spoken interaction between the user and the system. [Yankelovich 1996]

Spoken dialogue systems can be used heavily by some users and user groups. For vision-impaired users a speech interface may be essential for gaining access to information services. In a mobile context, users of productivity tools, such as email, may also be dependent on their speech-enabled access to information. Characteristic of such usage is that it occurs repeatedly, over a long period of time, and may use the features provided by the system in a versatile manner. Under such conditions, it is plausible that the users are willing to invest some effort into fully learning the possibilities offered by the speech-based service.

How, then, are speech-based systems introduced to new users? When speech is an additional modality in a system that comes with a manual, as is the case with the increasingly common voice control systems in automobiles [Heisterkamp 2001], the speech-based features can be described in the owner's manual. Even when users do not bother to read the manual [Carroll and Rosson 1987], they can discover the voice control possibilities through the graphical part of the car's computer interface [Pieraccini et al. 2004]. Unimodal, telephone-based spoken dialogue systems, on the other hand, need some auxiliary material to introduce them to the users. At least the telephone number and a brief description of what the service has to offer need to be provided to potential users. This is typically done through the web.

Service providers have created materials ranging from static web pages to multimedia presentations with audio examples of the interaction. In addition to introducing the service, users are also often provided with some instructions on how to use the system. Such a web-based tutorial can improve the user experience and users' perception of the system [Kamm, Litman, and Walker 1998]. Compared to guidance embedded into an interface, a comprehensive manual accessible through the web can result in more efficient interaction, since embedded guidance can make repeated interaction tedious.

Another approach to introducing new applications to users is the use of interactive tutoring. This is popular with applications that use graphical interface, particularly in video games, but it has been almost neglected in the case of speech-based applications. However, the tutorial type guidance can also be embedded into a dialogue system, e.g., as a specific guided mode, where new users receive extensive guidance to the systems and old users can use a more effective interface. A guided mode can make the system more transparent to users and thus help them, for instance, in knowing how to correct errors [Karsenty and Botherel 2005]. This kind of guidance can be extended by implementing a software tutor, a separate dialogue partner, which not only guides users but also monitors their interaction and makes sure that the users indeed learn to use the system. We have implemented such a tutor for an email reading application and studied its effect. It was found to reduce the amount of problems users have during the learning period [Hakulinen, Turunen, and Räihä 2006].

Here we follow-up our previous work on unimodal tutoring by studying a tutor that teaches spoken interaction using web-based graphical guidance. The multimedia tutor is connected to a spoken dialogue system so that a user can try out the system under the supervision of the tutor and receive guidance. The resulting multimodal [Oviatt 1999] or multimedia [Bearne, Jones, and Sapsford-Francis 1994] guidance has potential benefits. The visual presentation and GUI-based interaction can be superior to plain speech-based guidance by overcoming the transient and linear nature of speech and its rather low output rate.

We have implemented several versions of a multimedia tutor for a spoken dialogue application: a timetable system with a telephone-based speech interface. The technical challenges in implementing such a tutor that works in synchrony with the application have been discussed by Hakulinen, Turunen, and Salonen [2005a]. The tutors can be run as separate applications or embedded into other web-based guidance and marketing materials as Java applets. Four different concepts for the multimedia representation were developed and compared in earlier experiments [Hakulinen, Turunen, and Salonen 2005b]. The two most promising alternatives were chosen and implemented fully for the controlled experiments reported here.

The research question in this study is whether we can gain some benefits from extending graphical introduction material with interactive software tutor functionality. Can the interactive multimedia software tutoring indeed benefit users and how should the guidance be designed to be beneficial? We have conducted an experiment where the two versions of the interactive graphical software tutor are compared to a non-interactive web page version of the same material. All versions introduce the spoken dialogue system to users and guide them through an elementary scenario. The graphical tutors are connected to the dialogue system, and monitor users' interaction with the system and provide guidance as necessary. For example, after speech recognition rejections, guidance on how to speak to the system is given.

We collected data on users' interaction with the tutor and the dialogue system and users' attitudes towards the guidance materials and the system. An analysis of the measured data did not show significant differences in the task completion rates, but the most troublesome interactions occurred in the web guidance condition. The software tutor with more interaction possibilities was ranked highest in the subjective evaluations, while the other tutor was ranked the worst among the three conditions. Thus, the multimedia tutor can help, but only when designed properly. The graphical form used in the most interactive guidance helps users in understanding the functionality of the spoken dialogue system. The results point out the importance of constructing the guidance material in a manner that closely corresponds to the interaction model of the system: the interface is essentially a form-filling dialogue, and the highest ranked tutor is based on a graphical version of the form.

The paper continues with a description of the two tutors and the web-based guidance material and a discussion of their design rationale. The spoken dialogue system for which the guidance materials were built for is described as well. The experiment is then reported and its results presented. We close by discussing the implications of the work.

2. TWO INTERACTIVE TUTORS AND A WEB MANUAL FOR A TIMETABLE SYSTEM

We study the effects of interactive multimedia tutoring with two different tutors and a web manual. The tutors are graphical software applications run on a personal computer and they communicate with the spoken dialogue application running on a server. For comparison, a web manual has been constructed based on the tutors by removing all interactivity and arranging the information into a static document. All guidance material is in Finnish, as is the spoken dialogue system that the users are tutored in. In this paper the figures and examples of the guidance materials have been translated into English.

2.1 Busman Timetable System

The spoken language dialogue system that the tutors guide users on is called Busman. It is a research prototype of a telephone-based service for Tampere area public transport timetables [Turunen et al. 2005a], implemented in Java on top of the Jaspis framework [Turunen et al. 2005b]. Its users can ask for information on bus lines running between two locations, including information on departure times and routes. Typical utterances understood by the system include “Which line runs from University Hospital to the city center”, “When does the next bus to Hervanta leave”, and “When after six pm does a bus depart from Hervanta to university”. The system uses form-based dialogue management providing limited reasoning based on domain knowledge and dialogue history. Implicit confirmations are used extensively and mostly the interaction is user initiative. System initiative prompts are used for obtaining missing information and after repeated error situations.

The system uses the Finnish language. The speech recognizer is a commercial, large vocabulary recognition engine, Philips SDK with unisex Finnish acoustic models. The language model consists of about 1500 words, and is based on concept spotting. A

Finnish speech synthesizer, Mikropuhe by Timehouse with default male voice, is used for voice output. The system does not support barge-in, i.e., users cannot speak at the same time as the system but only when the system signals that it is listening. However, users can interrupt the system speech by pressing any key on the telephone keypad.

In addition to system initiative prompts, error messages etc., Busman features both short and rather exhaustive spoken help messages. Users can hear these messages by giving respective commands to the system. As an example, the response to a general help request says “This is the Busman system. You can query timetables for buses in Tampere. You can, for example, say ’which bus goes to Hervanta’, or ‘when does the next bus go to Hallila’. For comprehensive instructions, say ‘read instructions’.” The help functionality was available to participants during the experiment.

2.2 Tutor Design

The goal of the tutors is to introduce the Busman system to new users. In five to ten minutes, users will learn the functionality of the system and use it by following the instructions given by the tutor. The target group for the tutors is users who are new to the Busman system and possibly to spoken dialogue systems in general.

Since we already have experiences from speech-based tutoring and the new tutors are aimed to be a part of web-based information, the starting point for the tutor design was that the visual modality, which is not used by the Busman system, should be used. The only aural component in the tutors is a notification sound that is played via computer speakers when new tutoring material appears on screen. The sound is vital as it directs users’ attention from the application context to tutoring when necessary. The tutors include a visualization of the spoken interaction with comic book style speech balloons. Most importantly, this visualization shows users the speech recognition results as soon as they are available. The aim is to help the users in detecting and understanding speech recognition errors. System utterances are visualized as well, appearing just before the speech starts. This may help the users in learning to understand synthesized voice. Furthermore, the balloons displayed on the screen provide a short dialogue history for users, helping with the temporal nature of speech.

The tutors were presented to users as application windows as can be seen in Figures 1 and 2. In both tutors, the tutor window consists of a guidance area containing textual instructions, *Continue* and *Back* buttons to control the advance of tutoring, the visualization of spoken interaction with speech balloons and an outline of tutoring content in the bottom.

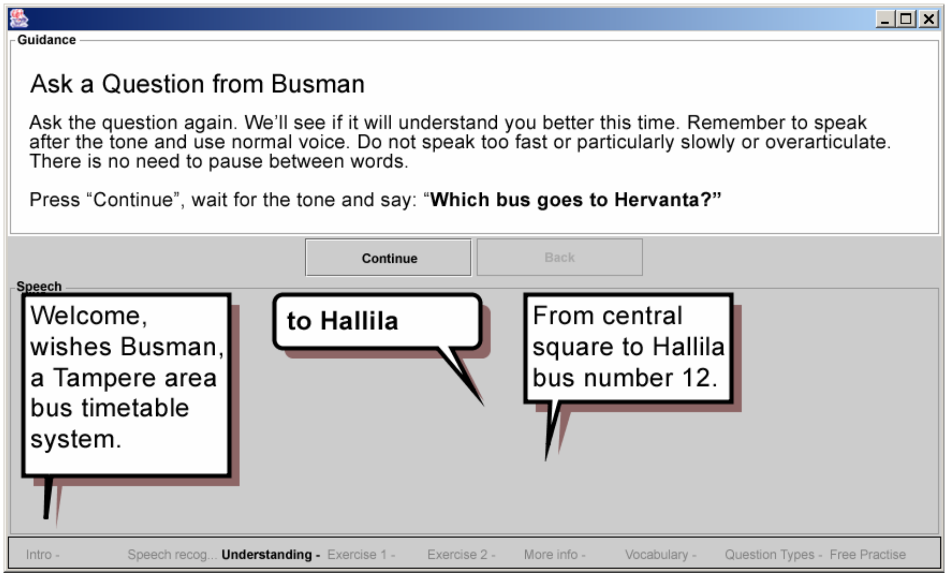
Guidance in both tutors is organized similarly into six segments, each consisting of one text screen. The users move between these by clicking the *Continue* and *Back* buttons. In addition to this text-based information, there is a hands-on exercise part in the middle of the tutoring. In the exercise, users are asked to try out Busman under the supervision of the tutor. This part consists of calling Busman and making three queries. In the end of tutoring, there is a possibility of free experimentation while the tutor is still active, i.e., visualizing the interaction and providing help in explicit error situations, but not directing the interaction. The last text segment before the free experimentation is a summary. Speech balloons are used to visualize spoken dialogue both during the hands-on exercise and free experimentation. They are also used to display an example dialogue before the exercise. The balloons use bold face font to emphasize keywords (the words and phrases the dialogue system actually use) in user utterances.

To support new users, the tutors guide the users step by step during the hands-on exercise. The users are told exactly what to say when they call the timetable system for the first time. The system usage is taught gradually with examples which the users try out themselves during the tutoring. The users control the pace of tutoring by pressing the *Continue* button. The tutors put the Busman system in a paused state as necessary so that the users can proceed at their own pace.

Since speech recognition errors are an inevitable part of speech-based interaction, error-related information is an important part of the tutoring. The tutors monitor the system, which it to display the recognition results in the balloons for users to see how the system has understood user input. During the hands-on exercise, the tutors first ask the users to give a specific input to the system and then monitor the speech recognition results for errors. The error analysis is based on the recognition result and the explicit request made to the user to give a specific input. By comparing the two on word and concept levels, the tutors can spot errors with certainty but the tutor cannot deduct their reason. The tutor does not make guesses; instead, it provides the user with guidance on how to remedy the situation. If the recognition results do not match the required input closely enough, help is given, and the user is asked to try again, simplifying the requested input if some information has already been given successfully. The help provided includes instructions on how to speak, such as to use normal voice and talk after a tone (see Figure 1 for an example). The users are asked to work on a single query for no more than three questions since we have noticed that people get frustrated by error loops at that point. By pointing out errors and providing relevant guidance, the tutors can help users in learning to detect, diagnose, and correct errors. In case of small word level errors, which still result in correct concept level outcome, the tutor points out the error and tells the user that it is not a significant one. For example, when a user has been asked to say “Which bus goes to Hervanta” and the recognition result is “Which to Hervanta”, the tutor tells the user “Busman heard your input and understood it correctly. It did not hear exactly ‘Which bus goes to Hervanta’ but it heard the important keywords and understood that you want to know the bus line number and you are going to Hervanta. Busman does not try to understand every word you say but it tries to find the important parts of your speech.”

2.3 Balloon Tutor

The first of the two tutors is the *Balloon tutor,* whose main feature is the visualization of spoken interaction between the timetable system and the user. Comic book style speech balloons are used: regular rectangles indicate utterances of Busman, rounded corners are used for utterances of the user. A snapshot of the Balloon tutor during the hands-on exercise part can be seen in Figure 1.



Guidance

Ask a Question from Busman

Ask the question again. We’ll see if it will understand you better this time. Remember to speak after the tone and use normal voice. Do not speak too fast or particularly slowly or overarticulate. There is no need to pause between words.

Press “Continue”, wait for the tone and say: “**Which bus goes to Hervanta?”**

Speech

Welcome, wishes Busman, a Tampere area bus timetable system.

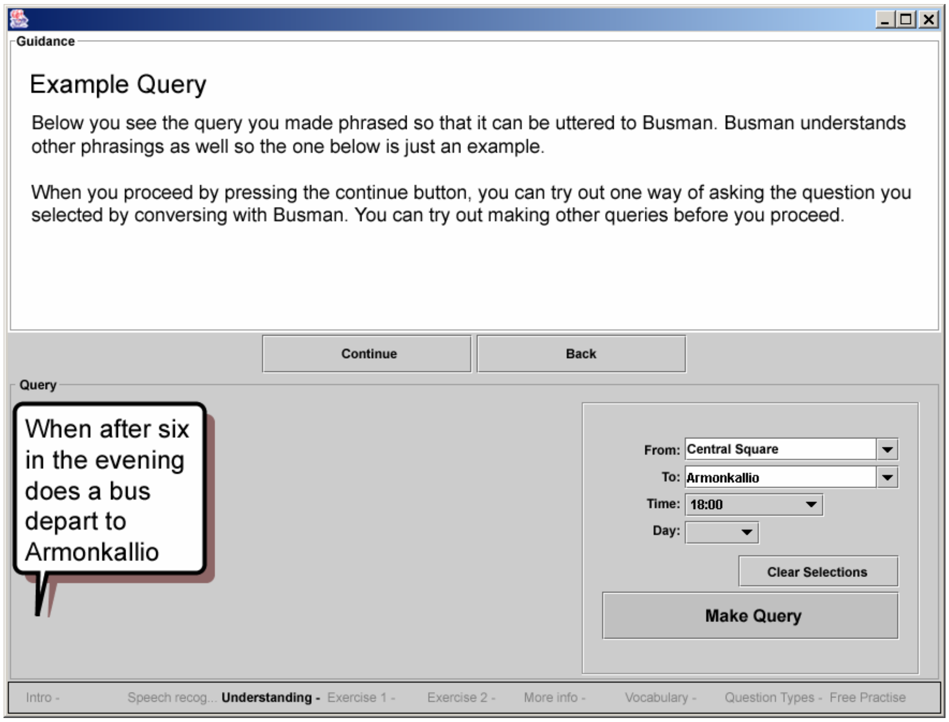
to Hallila

From central square to Hallila bus number 12.

Fig. 1: A screenshot of the Balloon tutor with a visualization of the spoken dialogue.

2.4 Form Tutor

The other tutor is called the *Form tutor*. It includes all the functionality of the Balloon tutor. In addition, it features a form consisting of graphical user interface components, which users can use to create queries that can be asked from the Busman system. The functionality found in the GUI form covers most of the features of the Busman system and therefore can be seen as a visual representation of the timetable system. The Form tutor, including the graphical interface, is shown in Figure 2. The Form tutor features the speech balloons and they are used to present the form-generated queries as well as to visualize the spoken interaction like in the Balloon tutor.



Guidance

Example Query

Below you see the query you made phrased so that it can be uttered to Busman. Busman understands other phrasings as well so the one below is just an example.

When you proceed by pressing the continue button, you can try out one way of asking the question you selected by conversing with Busman. You can try out making other queries before you proceed.

Query

When after six in the evening does a bus depart to Armonkallio

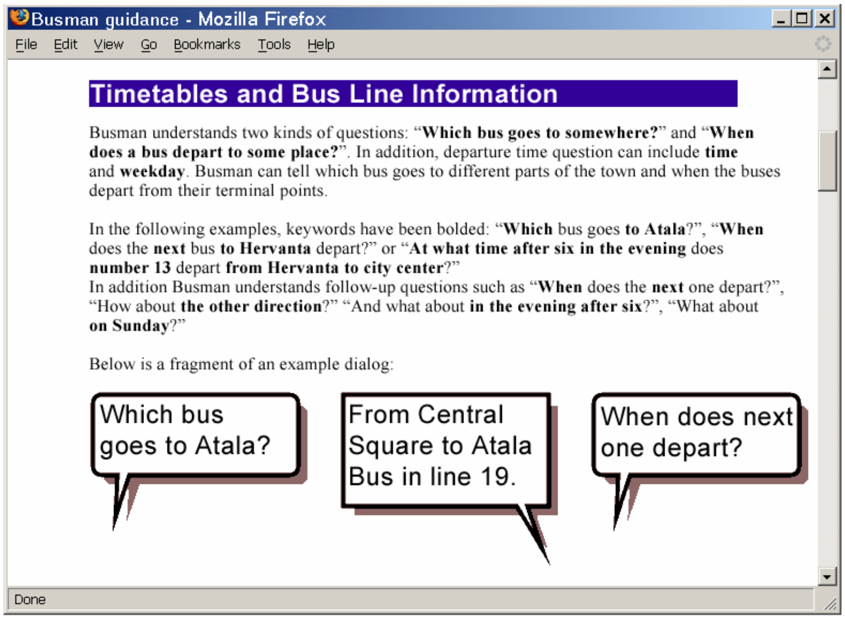
From: Central Square
To: Armonkallio
Time: 18:00
Day:

Fig. 2: A screenshot of the Form tutor.

Information provided by the two tutors is the same except for the form-specific information, i.e., the examples that the user can generate with the form. The queries the tutors ask the user to make are slightly different between the tutors. In the Balloon tutor, these requests are always the same while in the Form tutor they are based on a query created by the user with the form.

2.5 Web Manual

In addition to the two tutors, a web based version of the same material was created. It contains the same texts and graphics as the tutors as far as possible. However, the material is a single web page and therefore the only interaction users can have with it is to scroll the material up and down. The error related information, such as how to speak, which is presented in error situations in the tutors, is included in the web manual. The information has been slightly edited since error detection is the users’ task when the web manual is used. Part of the web manual can be seen in Figure 3.



Busman guidance - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Done

Timetables and Bus Line Information

Busman understands two kinds of questions: "**Which bus goes to somewhere?**" and "**When does a bus depart to some place?**". In addition, departure time question can include **time** and **weekday**. Busman can tell which bus goes to different parts of the town and when the buses depart from their terminal points.

In the following examples, keywords have been bolded: "**Which** bus goes **to Atala**?", "**When** does the **next** bus **to Hervanta** depart?" or "**At what time after six in the evening** does **number 13** depart **from Hervanta to city center**?"
In addition Busman understands follow-up questions such as "**When** does the **next** one depart?", "How about **the other direction**?" "And what about **in the evening after six**?", "What about **on Sunday**?"

Below is a fragment of an example dialog:

Which bus goes to Atala?
From Central Square to Atala Bus in line 19.
When does next one depart?

Fig. 3: Part of the web manual consisting of a single static HTML-page.

3. EXPERIMENT

The tutors and the web manual were used in an experiment to evaluate the effects of interactive guidance in introducing a spoken dialogue application to new users. Data on the effects of the use of the tutors was collected in the controlled experiment with 27 participants. They learned to use Busman under a controlled procedure. Interaction was recorded for later analysis and questionnaires were used to collect participants' opinions on guidance materials and the system.

There were three conditions, one for each guidance material, with 9 participants each. The conditions are called the *Web condition*, where the web manual was used as guidance, the *Balloon condition*, where the Balloon tutor was used, and the *Form condition*, where the Form tutor was used.

3.1 Procedure

The test had a between-subjects setting where each participant learned to use the timetable system in one of the guidance conditions. The participants were randomly assigned to the conditions.

Before the actual test, the participants approved that all calls to the system could be recorded and they filled in a background information questionnaire. The test consisted of a 15 minute learning period with the guidance and a 15 minute period for working with a set of 11 tasks without the guidance material. In the end of the experiment, users filled in two questionnaires where the timetable system and the guidance material used in the condition were evaluated.

The experiments took place in a regular office room where a fixed-line telephone was used to make calls to the Busman system and a desktop computer was used to access the guidance materials and fill in the questionnaires. The experiment conductor was present during the experiment but did not help users in their learning. He intervened only when it was time to move on in the experiment procedure or if technical problems arose.

3.2 Materials

A set of 34 questions known as SASSI (Subjective Assessment of Speech System Interfaces) [Hone and Graham 2000] was used to gather opinions on the Busman timetable system. A set of questions developed to evaluate usability and appeal of user interfaces by Hassenzahl et al. [2000] was used to gather opinions on the guidance. Both questionnaires used seven-item Likert-scale questions. An additional field for open comments was included at the end of both questionnaires. The questions were arranged in the questionnaires in a random order and the direction of scales in respect to their attitude towards the guidance and the timetable system varied between questions. The guidance questionnaire also included six additional Likert-scale questions on the length, amount, and consistency of guidance, resulting in a total of 28 Likert-scale questions. The questions were translated to Finnish for the study. When questions are mentioned in this paper, the original English language versions from the references are used. The questionnaires were presented to users as HTML-forms.

The tasks that the participants used Busman for were given on sheets of paper as maps with an arrow indicating the start and end locations of the trip and a clock face with am/pm next to it. One task description can be seen in Figure 4. In some tasks an abbreviation of Saturday or Sunday was also included, otherwise a weekday was to be assumed. The participants were asked to write down the bus line number for the requested route and a departure time that was near the given time. Place names were included in maps as text in the usual manner but the participants were allowed to also use their own knowledge of the town in selecting names for places. Task descriptions were given this way to minimize the amount of words given to the users. However, the first task was explained verbally to participants when the tasks were introduced to ensure that the task description was properly understood. In total there were 11 tasks starting from straightforward tasks with more complex tasks in the end. Complexities included such situations as lack of direct connections for the trip and departure times late at night when no more buses were running. All task descriptions were given to each participant at once and they were allowed to skip tasks and return to them as they pleased during the 15 minute period.

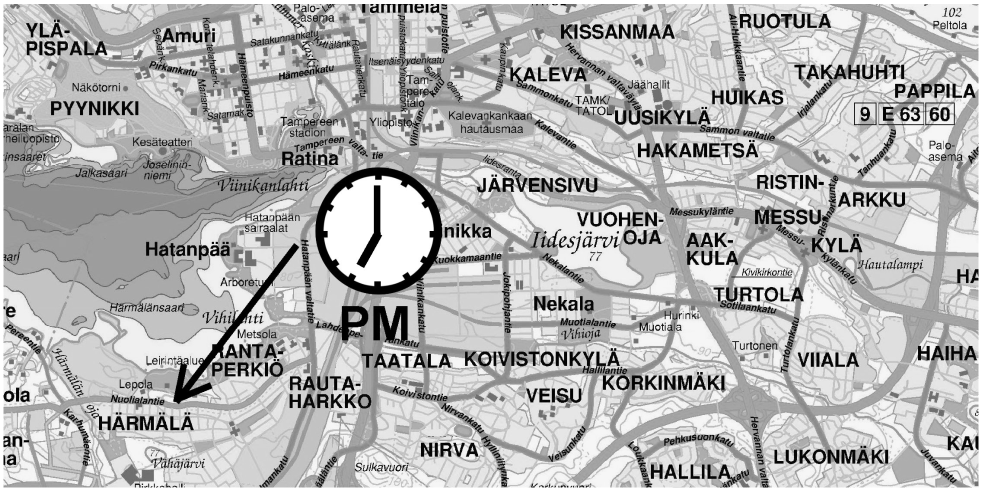


Fig. 4: A task description requiring participants to find a bus line from Hatanpään sairaala to Härmälä and a departure time for the line around 7 PM.

3.3 Participants

27 people without extensive experience in spoken dialogue systems or speech technology participated in the experiment. Their background information is summarized in Table I. The participants' familiarity with the town of Tampere and its public transport were inquired as they were relevant due to the application domain. There were no significant differences on background variables between the conditions. The participants were volunteers who replied to announcements placed on bulletin boards. They received a movie ticket for their participation.

Table I: Background variables.

Age	16-41, average 26.
Sex	10 male, 17 female.
Computer skills	Mostly common users, from inexperienced user to active hobbyist.
Years lived in Tampere	From 0 years to all their life.
Tampere area public transport use	From never used to regular users. Mostly occasional users.
Speech user interface experience	From never used to random usage.

3.4 Results

We have analyzed the task completion rates, which were similar in all conditions, the telephone calls, which reveal a wider variety of error rates in the Web condition, and questionnaires and general observations made during the experiments, which raise the Form tutor as the most highly ranked guidance type and provide some insights into differences between different kinds of users.

3.4.1 Task Completion. Participants' answers to the tasks were analyzed and scored to see if tutoring can teach the usage of the system successfully. There were no statistically significant differences on task completion between the conditions. On average, the participants were working with task number 8 when time was up and had successfully completed 5 to 6 tasks out of the 11 tasks within the time limit in every condition. Some of the earlier tasks were also left blank or given incomplete answers, as can be seen in Table II. In the table, scores given for participants' answers have been summed per condition and task. Scores were given so that completely wrong answers and missing answers received 0, half point was given in cases where only part of the answers was correct (e.g., only bus line but no time) and correct answers resulted in 1 point. None of the participants could give an answer to the final task within the time limit. Users' success with the tasks and the similar task completion rates tell that the tutors can teach the usage just as well as the web manual and the interactivity does not hinder learning. However, the benefits of interactivity cannot be seen in the task completion rates.

Table II: Sum of task scores per task and condition.

Task	1	2	3	4	5	6	7	8	9	10	11
Web condition	6,5	6	6	7,5	6,5	5,5	5,5	4	2	2	0
Balloon condition	6	6	4,5	8,5	6	6	6	3	1	1,5	0
Form condition	6,5	5	6,5	8,5	8,5	6,5	3,5	3	3	2	0

3.4.2 Interaction with the System The participants' interaction with the Busman system both during the training period and during the tasks was recorded and analyzed. The metrics show what kind of problems the participants had and what the effects of tutoring were. Most importantly, users' interaction with tutors seems to be more consistent while some users of a static manual do just fine and others have serious problems.

Speech recognition rates were calculated in the form of concept recognition rates, i.e., as the percentage of the correctly recognized concepts in all utterances. An average concept recognition rate over all participants was 79%. Concepts in the Busman domain are such things as departure place, destination, departure time, help request etc. There were some interspeaker differences in the concept recognition rates, but between the conditions the recognition rates were very consistent as can be seen in Table III.

Table III: Concept recognition rates for the conditions.

	Average	Min.	Max.
Web condition	77 %	64 %	88 %
Balloon condition	79 %	70 %	86 %
Form condition	79 %	74 %	88 %

While there were no statistically significant differences in the error rates between the conditions, the variances of utterance level error rates (i.e., percentage of utterances that did not result in correct system response) between the three conditions were significantly different (Bartlett test of homogeneity of variances, df = 2, $p < 0.05$). The Web condition had the highest variance in error rates while the Balloon condition had the lowest. The difference in variances can be seen in Figure 5 where rates of total recognition error free utterances during the entire experiment have been plotted for each participant. The large variance in the Web condition can be seen especially compared to the Balloon condition.

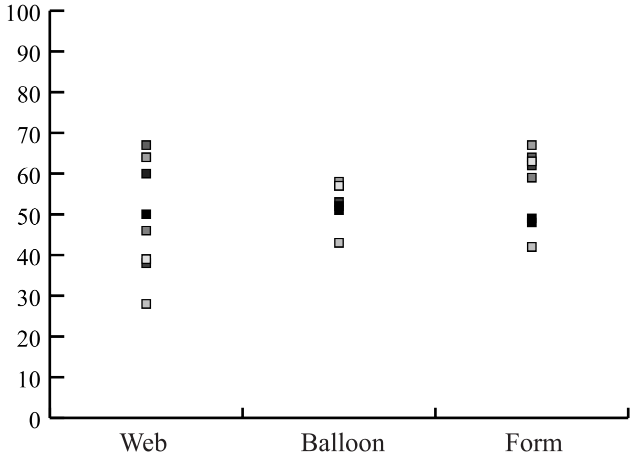


Fig. 5: Total rates of recognition error-free input per participant.

When the training part, i.e., the direct effect of the condition is removed, and only the interaction during the tasks is considered, the error rate distributions become more similar and the difference of variances is no more statistically significant.

In addition to speech recognition errors, voice activity detection (VAD) errors were analyzed. Most of these were cases where end of speech was detected, usually due to a small pause, when a user was still planning to continue the utterance. In such cases, the start of the utterance was often recognized correctly but the user had to repeat in the next utterance the concepts included in the part lost by the system. In general, users were able to detect the voice activity detector errors in all conditions since tones were used to signal both start and end of the period when the system was listening. There were also errors where VAD did not react to user's speech at all and the whole utterance had to be repeated. While the Web condition had more VAD related errors than the other conditions, as can be seen in Table IV, the difference is not statistically significant. Speech recognition errors and VAD errors are independent, i.e., any user utterance could contain no errors, only speech recognition errors, only VAD errors or both.

Table IV: Voice activity detector (VAD) errors as percentage of user utterances affected by an error.

	Average	Min.	Max.
Web condition	26 %	10 %	41 %
Balloon condition	18 %	11 %	29 %
Form condition	18 %	5 %	32 %

Percentages of other types of problems, interruptions of system prompts by the participants using telephone keys, and the total numbers of utterances for the training and task periods can be seen in Table V. The significant differences were the numbers of utterances and the numbers of interruptions during the training ($p < 0.05$ and $p < 0.0005$, respectively, on Kruskal-Wallis one way analysis of variance, df = 2). These differences are directly dependant on the controlled variable, i.e., the training material used. The difference in the amount of interruptions was also significant during the entire experiment ($p < 0.05$) but this can be attributed to the very large difference during the training period.

Table V: Average percentages of utterances spoken at a time when the system was not listening, containing Out-Of-Vocabulary concepts, and followed by an interruption of the system response, and the number of utterances per condition.

	Web		Balloon		Form	
	Training	Task	Training	Task	Training	Task
Wrong time %	2,0 %	1,0 %	4,0 %	3,2 %	6,6 %	1,8 %
OOV %	2,5 %	5,6 %	2,4 %	6,9 %	5,3 %	6,2 %
Interruptions	0,0 %	2,5 %	6,1 %	2,2 %	1,7 %	6,0 %
# of utterance	11	41	23	44	14	41

3.4.3 Questionnaires Participants' subjective evaluations on the guidance materials and the spoken dialogue system were collected using questionnaires. These evaluations provide information on which guidance material the participants liked most as well as some insights into the effect of computer skills and some other background variables on the evaluations.

The guidance evaluation questionnaire resulted in different overall evaluations for the guidance materials. The differences are highly significant (Friedman rank sum test (of evaluation medians), df = 2, $p < 0.001$). Rank sums (higher value – better evaluation) were 55.5 for the Web condition, 39.5 for the Balloon condition, and 73.0 for the Form condition. There were no statistically significant differences between the conditions within single guidance evaluation questions.

There was no significant difference between the conditions on the SASSI evaluation of the Busman system. However, participants' backgrounds correlate with some evaluations, raising the question of differences between different groups of users. Computer skills is a variable that highly significantly correlated (Pearson's product-moment correlation df = 25, p < 0.01) with answers to several questions: "The system is pleasant", "The system is friendly", "The interaction with the system is irritating", "The interaction with the system is frustrating" and "The system responds too slowly". In all cases more experienced computer users considered the timetable system worse, i.e., less pleasant and more irritating. This is understandable as people who are more knowledgeable of computers may have higher expectations of such systems. Their expectations may also be based more on how desktop computers behave. Speech user interface experience correlates also with computer skills (Pearson's product-moment correlation, df = 25, p < 0.05). However, computer skills did not correlate with error levels or task completion rates. Furthermore, the correlations of computer skills were only with system evaluations. There was no significant correlation with the guidance evaluations, which suggests that the tutors, while not equally necessary to, were equally accepted by the different types of users. There was also a difference between male and female participants' evaluations on the question "The system is easy to use". Women agreed more that the system is easy to use. This might be because female users reported lower computer skills than male users. Women had also a slightly lower total error rate in their interaction with Busman, but the difference is not significant.

In guidance questions age correlated (Pearson's product-moment correlation, df = 25, $p < 0.001$) negatively with answers to the question "Guidance was too long", i.e., younger participants considered the guidance too long more often than older ones, suggesting that older users are willing to study more exhaustive guidance material than young users.

4. DISCUSSION

Spoken dialogue systems need to be introduced to new users somehow. Rarely does a potential user receive only a phone number to call to. They usually read some marketing material and other introductory information before they decide to call a system. This material is commonly available in the web, and in addition to marketing material, or as a part of it, the information commonly contains some form of guidance on how to use the system.

In this study, we compared different guidance materials that can be embedded into other web-based material to teach to use of a spoken dialogue system. The materials are aimed to users new to the dialogue system and possibly to spoken dialogue systems in general. Two interactive graphical software tutors and a static web manual were used to teach how to use the spoken dialogue system. The aim was to see if we can gain some benefits by using interactive multimedia software tutoring, and how should such tutoring be designed.

The results indicate that interactive tutoring helps especially those people, who would have most problems learning the use with static guidance materials. While some users can learn to use a system just fine with just a static manual or even without any guidance, others have many problems in learning the style of interaction required in human-computer spoken dialogue. Unlike static guidance, tutors were able to take care of all users. This can be seen in the variations of error rates between the different guidance conditions, i.e., some users in the static guidance condition had much more problems than others while tutors provided learning results that are more consistent. This supports our experiences with speech-only tutoring [Hakulinen, Turunen and Räihä 2006], where the interactive tutor could reduce the number of problems users had during the learning period, making sure that everybody could learn the system smoothly.

We carried out a small follow-up experiment with the multimedia tutors with an additional condition where no guidance material was available. The results suggest that variance in error rates will be even higher when no guidance is available. One user in that condition received the best task completion speed of all participants by learning with just trial and error. Another user in the same group gave up during the learning period. She was given an opportunity to try the system with the Balloon tutor and could proceed successfully. It is also worth mentioning, that especially those, who felt more insecure on using the system, reported that they felt comfortable when they received support from the tutor in the beginning. Tutoring can support users who could not learn the system otherwise, but not all users should be forced to use one.

Since speech is considered a natural interaction method, the success of speech-based systems depends heavily on matching users' expectations. If users assume that a system can understand spontaneous speech like a human, they are very likely to be disappointed. On the other hand, too low expectations can make users ignore the system, or use only a part of its functionality. Tutors support learning by making sure that they use the system correctly. Overall, in speech-based human-computer interaction, error recognition and correction are very important due to the probabilistic nature of speech recognition. Error correction methods and capabilities of spoken dialogue systems are much more limited than those of human-human interaction. Therefore, errors are important part of guidance material as well. By giving explicit instructions to users on what to say, the tested tutors could detect speech recognition problems, which the spoken dialogue system itself is not capable of spotting. In addition, the tutors enabled users to spot recognition errors from the visualization of speech recognition results. Other problems, like speaking so softly that the voice activity detector did not detect any speech, or speaking when the system was not listening, were corrected by tutors very efficiently since the error condition could be analyzed well by the tutor. When interactive guidance is not used, great care must be taken to produce supporting static guidance to helps users to learn error detection and recovery. Naturally, the spoken dialogue system must have sufficient error handling functionality to be taught to users.

Concrete visual example dialogues were a successful design decision, both in the static web page and as visualizations of the spoken dialogue in the tutors. The participants liked the example dialogues presented as speech balloons. They worked well also in the static web manual where participants commonly scrolled the page so that an example dialogue was visible on screen when they called the system for the first time. The solution of presenting the keywords understood by system with boldface in the balloons, while being designed for the real-time visualization, received a positive comment also in the Web condition, where words in static text were bolded. This kind of examples of basic interaction with a system are very easy to produce and can support insecure users when they try to use a system for the first time.

Participants' evaluations of the guidance materials raise the most interactive guidance, Form tutor, significantly above the Balloon tutor and above the static guidance as well. A form with a graphical user interface in the Form tutor, presenting the same functionality as the spoken dialogue system, was the only difference between the two interactive tutors. The benefit of a graphical form is in line with a finding by Terken and teRiele [2001] that a multimodal interface with a graphical query interface provided a mental model that was useful with a speech only interface. The results of our study show that a similar approach featured in the Form tutor was successful and well received by users. The fact that users can transfer skills from a graphical interface to spoken interaction has implications not only in designing introductory material for speech interfaces, but also to developers of multimodal systems. The findings also show that the design and features of interactive tutoring are important; the guidance material should closely correspond to the interaction model of the system.

REFERENCES

BEARNE, M., JONES, S., AND SAPSFORD-FRANCIS, J. 1994. Towards usability guidelines for multimedia systems. In *Proceedings of the second ACM international conference on Multimedia*, San Francisco, CA, USA, October 1994, ACM Press, 105-110.

CARROLL, J. M. AND ROSSON, M. B. Paradox of the Active User. In *Interfacing Thought: Cognitive Aspects of Human-Computer Interaction*. Cambridge, J. M. CARROLL, Ed., MIT Press, MA, 1987.

COX, R. V., KAMM, C. A., RABINER, L. R., SCHROETER, J., AND WILPON, J. G. 2000. Speech and Language Processing for Next-Millenium Communications Services. *Proceedings of the IEEE*, 88, 8, 1314-1337.

HAKULINEN, J. TURUNEN, M., AND RÄIHÄ K.-J. 2006. Evaluation of Software Tutoring for a Speech Interface. *International Journal of Speech Technology*, 8, 3, 283-293.

HAKULINEN, J., TURUNEN, M., AND SALONEN, E.-P. 2005a. Visualization of Spoken Dialogue Systems for Demonstration, Debugging and Tutoring. In *Proceedings of Interspeech'2005 - Eurospeech — 9th European Conference on Speech Communication and Technology,* Lisbon, Portugal, September *2005*, ISCA, 853-856.

HAKULINEN, J., TURUNEN, M., AND SALONEN, E.-P. 2005b. Software Tutors for Dialogue Systems. In *Proceedings of Text, Speech and Dialogue*, LNAI 3658, MATOUSEK, V., MAUTNER, P., AND PAVELKA, T. Eds. Springer, 412-419.

HASSENZAHL, M., PLATZ, A., BURMESTER, M., AND LEHNER, K. 2000. Hedonic and ergonomic quality aspects determine a software's appeal. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, The Hague, The Netherlands, April 2000, ACM Press, 201-208.

HEISTERKAMP, P. 2001. Linguatronic product-level speech system for Mercedes Benz cars. In *Proceedings of the 1st International Conference on Human Language Technology Research, HLT '01,* San Diego, CA, USA, March 2001. ACL, 1–2.

HONE, K., AND GRAHAM, R. 2000. Towards a Tool for the Subjective Assessment of Speech System Interfaces (SASSI), *Natural Language Engineering, Best Practice in Spoken Language Dialogue System Engineering,* Special Issue, 6, 3 & 4, September 2000.

KAMM, C., LITMAN, D., AND WALKER, M. A. 1998. From Novice to Expert: The Effect of Tutorials on User Expertise with Spoken Dialogue Systems. In *Proceedings of the 5th International Conference on Spoken Language Processing,* Sydney, Australia, November-December 1998. MANNELL, R.H., AND ROBERT-RIBES, J. Eds. ASSTA, 1211-1214.

KARSENTY, L., AND BOTHEREL, V., 2005 Transparency strategies to help users handle system errors. *Speech Communication*, 45, pp. 305–324.

OVIATT, S. L. 1999. Ten myths of multimodal interaction. *Communications of the ACM*, 42, 11, 74–81.

PIERACCINI, R., DAYANIDHI, K., BLOOM, J., DAHAN, J.-G., PHILLIPS, M., GOODMAN, B. R., AND PRASAD, K. V. 2004. Multimodal conversational systems for automobiles, *Communications of the ACM*, 47, 1 Multimodal interfaces that flex, adapt, and persist.

TERKEN, J., AND TE RIELE, S. 2001. Supporting the Construction of a User Model in Speech-only Interfaces by Adding Multi-modality. In *EUROSPEECH 2001 Scandinavia,, 7th European Conference on Speech Communication and Technology, 2nd INTERSPEECH Event,* Aalborg, Denmark, September 2001, DALSGAARD, P., LINDBERG, B., BENNER, H., AND TAN, Z-H. Eds. ISCA, 2177-2180.

TURUNEN, M., HAKULINEN, J., SALONEN, E-P., KAINULAINEN, A., AND HELIN, L. 2005a. Spoken and Multimodal Bus Timetable Systems: Design, Development and Evaluation. In *Proceedings of 10th International Conference on Speech and Computer*, Patras, Greece, October 2005, 389-392.

TURUNEN, M., HAKULINEN, J., RÄIHÄ, K-J., SALONEN, E-P., KAINULAINEN, A , AND PRUSI, P. 2005b. An architecture and applications for speech-based accessibility systems. *IBM Systems Journal*, 44, 3, 485-504.

YANKELOVICH, N. 1998. How Do Users Know What to Say? *Interactions*, 3, 6, 32-43.