SPECOM 1362
4 November 2004 Disk Used
ARTICLE IN PRESS
No. of Pages 16, DTD = 5.0.1

ELSEVIER

Available online at www.sciencedirect.com
SCIENCE DIRECT®

Speech Communication xxx (2004) xxx–xxx

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Effects of displayless navigational interfaces on user prosodics

Julie Baca [a,*,1], Joseph Picone [b,2]

[a] *US Army Corps of Engineers, Waterways Experiment Station, Vicksburg, MS 39180, United States*
[b] *Department of Electrical and Computer Engineering, Institute for Signal and Information Processing, Mississippi State, MS 39762, United States*

Received 1 August 2003; received in revised form 1 July 2004; accepted 28 September 2004

Abstract

Displayless interface technology provides speech-based access to computer applications for which visual access is not possible. These applications are increasingly prevalent, especially in situations requiring mobility, such as navigational applications. To ensure the successful deployment of this technology however, many human factors issues must be addressed. In particular, its nonvisual nature requires verbal presentation of spatial data. Prosodics, or nonverbal aspects, of human speech have been established as an indicator of cognitive stress. In this paper, we examine the assumption that the cognitive burden placed on the user by displayless access to spatial data would significantly alter the prosodics of the user's speech.

Results were gathered through experiments in which user interactions with a prototype speech-based navigational system were recorded, post-processed, and analyzed for prosodic content. Subjects participated in two sessions, one using a speech-based, displayless interface, and a second using a multimodal interface that included a visual–tactile map display. Results showed strong evidence of significant changes in subjects' prosodic features when using a displayless versus a multimodal navigational interface for all categories of subjects. Insights gained from this work can be used to improve the design of the user interface for such applications. Also, results of this work can be used to refine the selection of acoustic cues used as predictors in prosodic pattern detection algorithms for these types of applications.

Keywords: Prosodics; Displayless; Multimodal

* Corresponding author. Address: Center for Advanced Vehicular Systems, Mississippi State University, 200 Research Blvd, Starville, MS 39759, United States. Tel.: +662 325 5442/004; fax: +662 325 5543/7300.
E-mail addresses: baca@cse.msstate.edu, baca@cavs.msstate.edu (J. Baca), picone@isip.msstate.edu (J. Picone).
[1] Center for Advanced Vehicular Systems, Engineering Research Center, P.O. Box 9627, Mississippi State, MS 39762, United States.
[2] Tel.: +662 325 3149; fax: +662 325 2298.

0167-6393/$ - see front matter
doi:10.1016/j.specom.2004.09.006

1. Introduction

The graphical user interface (GUI) created a fundamental shift in the nature of human–computer interactions from a style that was strongly text-based to one that is predominantly visual. Ironically, concurrent to the growth in popularity of the GUI, research and development of displayless interface technology has also advanced. Displayless interface technology provides speech-only access for applications in which the use of a visual interface is not possible or is greatly restricted, such as those requiring mobility or the use of a cellular telephone. Often this technology must verbally present data that is either spatial in nature, such as geographical maps, or data that is presented through a visuospatial display metaphor, i.e., a GUI. Results of research presented in this paper strongly support the assumption that presentation of spatial data through a strictly verbal interface modality increases the cognitive load for the user. Results were gathered through experiments in which subjects used a displayless navigational interface for the US Army Corps of Engineers Waterways Experiment Station (Baca, 1998). Subjects used the program *WES Travel* to plan routes around the station through speech-based as well as multimodal interaction.

A navigational displayless interface was chosen for testing since, despite its limitations, speech provides a desirable alternative for many applications in which spatial data must be presented nonvisually, particularly those requiring mobility. For example, systems described in (Baca et al., 2003; Buhler et al., 2002; Pellom et al., 2001) allow drivers to query for information regarding geographical routes from one location to another. The use of similar technology in a mobile navigational aid for visually impaired travelers in unfamiliar environments was investigated by Loomis et al. (1994). Indeed, the latter category of users are uniquely affected by the quality of displayless interface technology.

For all users of this technology, however, widespread use will require addressing many issues in the realm of human–computer interaction. This study investigated one issue in particular, speaker prosodics. Previous research, reviewed by Scherer (1981), examined the impact of psychological and cognitive burdens on the prosodics of human speech, e.g., fundamental frequency (F0), speaking rate, and the length and location of pauses. More recent work conducted by Scherer et al. (2002) found significant effects of cognitive load due to task engagement on prosodic features including, speaking rate, mean F0 and energy. The study entailed recording the speech of subjects performing a logical reasoning task requiring cognitive planning. The task was presented visually to subjects on a computer screen with no speech output. The research presented in this paper extends the study of Scherer et al. (2002) by examining the possible increased cognitive load due to performing a similar type task, spatial planning, with only verbal description and no visual presentation on the screen, and the effects of this load on the prosodics of the user's speech. A better understanding of this issue could contribute to the development of more robust interfaces using better prosodic pattern detection for applications requiring displayless access to spatial data.

As noted by Noth et al. (2000), prosody plays a significant role in disambiguation in human–human communication. The nature of displayless interactions more closely resembles this type of communication since computer speech functions in the role of the human. Analogous to how pauses, intonation, and register of a human speaker convey meaning to the human listener, these characteristics of computer speech convey meaning to the user. Similarly, prosodic information contained in the user's speech, such as the change in duration of phonemes or the presence of embedded silences, can also convey meaning. Consider this sentence in a navigational task

"Where can I find CH, IT, and EL?" versus
"Where can I find CHIT, and EL?"

where CH is commonly used to abbreviate the Coastal and Hydraulics Laboratory, IT is commonly used to refer to both a separate laboratory, Information Technology (IT) Laboratory, as well as the IT department within the Coastal and Hydraulics Laboratory, and finally, EL denotes the Environmental Laboratory. The two sentences differ prosodically; when spoken, the first sentence contains an embedded pause between the character combinations, "CH" and "IT". The presence or absence of a pause conveys two very different meanings for the two sentences. However, the results reviewed in (Scherer, 1981) and the findings of Scherer et al. (2002) indicate that both hesitation pauses and speaking rates tend to increase in tasks requiring cognitive planning, rendering either of these cues alone less accurate predictors of phrase boundaries. Therefore, in the example sentence, a pause between "CH" and "IT" may indicate cognitive load, not a conscious attempt to delineate these two entities.

The previous example illustrates how knowledge gained from investigating the effects of cognitive load on prosodics can be used to improve prosodic pattern detection algorithms for applications that require cognitive planning, such as displayless navigational systems. Prosodic information has been used to reduce syntactic ambiguity in sentence parsing (Price et al., 1991) as well as to detect phrase boundaries (Wightman and Ostendorf, 1994). Wightman and Ostendorf (1994) discussed the limitations of algorithms using limited acoustic cues such as F0 or other single features. They proposed that a combination of acoustic cues, including pauses and other durational features, should be used for more robust prosodic pattern detection. A correlation between the additional cognitive load induced by displayless navigational interfaces and changes in the prosodics of the user's speech lends support to this argument since this variability would render single cues less robust predictors.

Algorithms to detect prosodic patterns in speech have addressed several problems, including phrase structure recognition relying on the use of F0 contour analysis (Huber, 1989; Nakai et al., 1994; Okawa et al., 1993), tone recognition to classify boundary tones and detect yes/no questions from F0 contours (Daly and Zue, 1990; Waibel, 1988), and stress detection algorithms to detect the relative prominence of a syllable (Campbell, 1992; Chen and Withgott, 1992). Many of these approaches used only limited acoustic cues. The algorithm developed by Wightman and Ostendorf (1994) used multiple prosodic cues, including pauses, boundary tones, and speaking rate changes to detect phrase boundaries. It also worked with the output of a speech recognizer rather than the actual speech signal. The algorithm was tested on two corpora of professionally read speech and achieved agreement between automatically detected and hand-labeled results comparable to human inter-labeling agreement.

More recent research using prosody in speech understanding in the VERBMOBIL project used both the output of a speech recognizer and the speech signal (Noth et al., 2000). In addition, this research analyzed spontaneous speech collected from human–human dialogues. This approach yielded best results, e.g., absolute recognition word accuracies of 91% and 92% when multiple features, including duration, F0, energy, and speaking rate, were used. Parsing time was also reduced by 92%.

To reiterate, increased cognitive loading during interactions with displayless navigational interfaces may cause the user to alter his or her prosodics; further, changes in the user's prosodics could significantly affect the performance of prosodic pattern detection algorithms for these applications. This is particularly relevant for current dialog systems providing navigational information, such as (Baca et al., 2003; Buhler et al., 2002; Pellom et al., 2001). The remainder of this paper is organized as follows: Section 2 describes the experimental methods used to test fundamental assumptions of the research; Section 3 describes results, and Section 4 presents conclusions and potential areas for future work.

2. Experimental methodology

Testing the assumption that the prosodics of the user's speech while interacting with a displayless navigational system would differ significantly from that produced while interacting with a multimodal navigational system required analyzing recordings of user speech interactions with a prototype displayless interface to a map database of the USACE WES. A map of the area is included in Fig. 1. Subjects participated in a single experiment, consisting of two sessions. During each session,

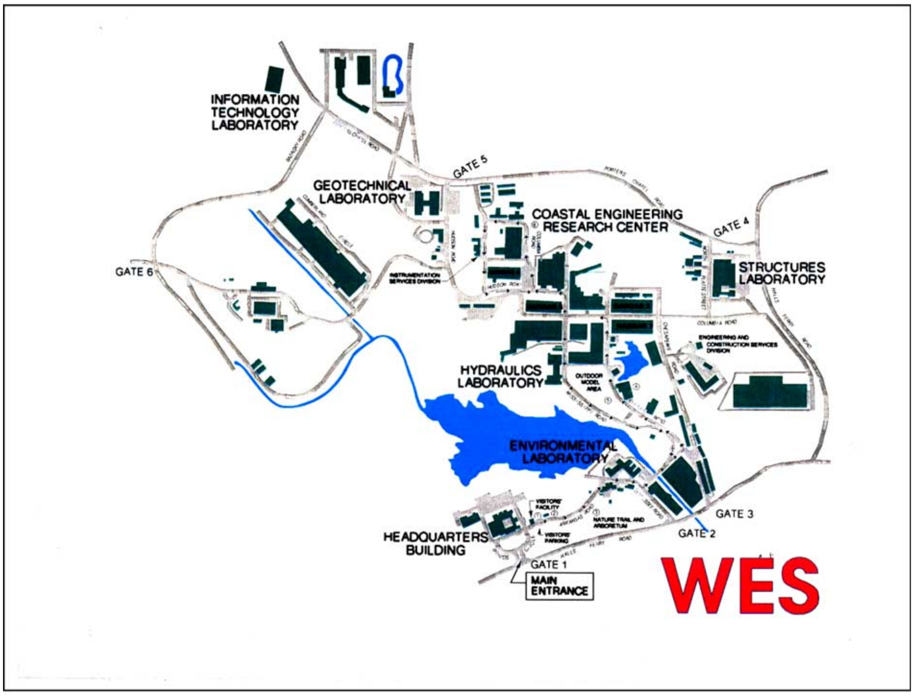

Fig. 1. WES map.

subjects performed a series of increasingly complex navigational tasks.

The assumptions regarding cognitive load were deemed applicable to all users, irrespective of visual acuity. Details of results for subjects with visual impairments are given in (Baca, 1998). This paper also includes detailed results for sighted subjects. In the first session, all subjects used only a speech interface to perform the tasks; in the second session, sighted subjects used a multimodal audio-graphical display, while subjects with visual impairments used an audio-tactile display. User speech was recorded during each session, post-processed for prosodic content and statistically analyzed for differences in prosodics between the two sessions. The following sections describe three components of the experimental methodology: Section 2.1 reviews key aspects of the speech-multimodal prototype used in the experiments; Section 2.2 discusses critical issues in subject selection, and Section 2.3 describes the tasks performed by subjects in the experiments.

2.1. A prototype travel information system

The prototype used in the experiment, WES Travel, consults the map database to give spoken instructions to visitors attempting to locate areas of interest. Visitors can query for specific instructions or ask the program to compute a driving route from one location to another. During the experiments, subjects were asked to assume the role of first-time visitors to the station and use the program for assistance in getting from one location on the station to another with the stipulation that the route they planned be safe for pedestrians. Information relevant to pedestrians, such as sidewalks and crosswalks, was contained in the map database as well as that relevant to both drivers and pedestrians, e.g., traffic and road construction. After listening to a verbal description of the overall station layout, subjects were given a starting point and a destination for each task and then asked to use the program to determine an optimal walking path to the destination.

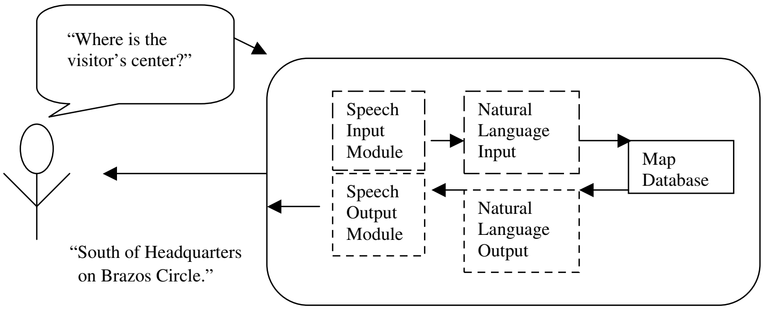

Fig. 2. Prototype travel information system.

In the first session, subjects used a speech-only interface. All interactions between the user and the system were conducted through speech, as shown in Fig. 2. The speech input module used an automatic speech recognition (ASR) engine. The rationale for the use of ASR rather than a Wizard-of-Oz (WOZ) approach was based on findings from research conducted using the Air Travel Information System (ATIS), a displayless application providing information to travelers (Godfrey and Doddington, 1990). Research demonstrated that as the word error rate (WER) reaches approximately 10% or lower, it is highly correlated with the language understanding error rate (Bayer et al., 1995), the latter of which directly impacts a user of a navigational application, which functions as an information querying rather than dictation style program. Further, Dahl et al. (1994) argue that using ASR versus a WOZ yields more realistic data for analysis since it obtains data from subjects who are actually speaking to a computer. Therefore, the speech input module provided speaker-independent recognition of continuous speech using the Entropic HTK ASR engine (Woodland et al., 1994), trained on the DARPA Wall Street Journal (WSJ) corpus (Paul and Baker, 1992) with a WER of 8.1% and a real-time factor of 2XRT running on a 100 MHz processor. The acoustic conditions in the experiments were carefully controlled so that the WSJ models would provide an appropriate match to the speech data collected. The fielded system used a vocabulary of approximately 6000 words, including the 5000 WSJ vocabulary with approximately 1000 business and other domain specific words interpolated with the WSJ using a back-off N-gram model. The fielded system performed with an absolute WER of 10.2% for the ASR and a semantic error rate of 13.9%. In addition, to further reduce any impact of recognition or understanding errors on the results of the investigation, a minimal error-handling strategy, as recommended in (Kamm, 1994), was used. Requests were confirmed only when the consequences of an error could cause significant inconvenience to the user. The NL parser uses a semantic grammar and limited contextual knowledge of previous queries to parse and translate requests into database queries. This allows input of freely formed natural language queries to obtain information such as, "What's the road like from here to the visitor's center?" or "Is there a sidewalk on this road and is traffic heavy here?"

Avoiding auditory overload presented a significant issue in the design of the speech output module due to the spatial nature of the data presented. The research presupposed an increase in the user's cognitive load due to verbal presentation of such data; however, this could only be tested with accuracy if auditory overload were minimized. Measures taken to address this included reducing the use of auditory lists and speaking directions in brief segments which the user could easily request to be repeated.

Another consideration for the speech output module concerned the presentation of directional information. Previous research indicated that people vary widely in their understanding and use of compass directions, i.e., north, south, east, west (Kozlowski and Bryant, 1977; Thorndyke and Stasz, 1980) and thus prefer multiple categories of directional information when receiving directions. Therefore, the program combines compass directions, commonly used directional language, such as "left", "right", "behind", and "ahead", as well as prominent stationary landmarks. This reduces the ambiguity of instructions, but increases the amount of information spoken to the user and thus, the potential for auditory overload. To minimize this, the program gives orientation in several short segments, each repeatable by pressing a key. Examples of such instructions at the onset of a route are given in Section 2.3.

In the second session, subjects used an interactive touch screen display of a map of the station in addition to speech. Key areas were visually and tactilely highlighted on the map for selection. Users could touch the selectable areas on the map and hear short descriptions of the areas as well as query through speech, as in the first session.

For the multimodal interface, design of the graphical interface adhered to the design goals of offering completeness while maintaining simplicity. These objectives motivated the selection of the map for the display designed by a graphic artist for station visitors, rather than a detailed drawing produced from the original database for WES engineers and maintenance personnel. This provided a more intuitive view for users unfamiliar with the station. Design of the tactile display adhered to similar design goals as that of the graphical; however since it could not provide the same level of detail meaningfully, design guidelines by Barth (1983) for creating tactile maps were followed. Further details of the audio and tactile display as well as other features of the prototype are given in (Baca, 1998).

2.2. Subject selection

Selection criteria applied to all subjects included age, education, and amount of previous computer experience. All subjects were required to be 18 years of age or older and possess the equivalent of at least a high school education, i.e., high school diploma or General Equivalency Diploma. Also, all subjects were required to be current users of computer software, performing some type of task regularly, i.e., at least weekly or monthly, with no restrictions on the nature of the software or task. This ensured a baseline of experience in computer usage. Finally, all subjects were required to have no previous knowledge of the physical layout of the WES.

While users with visual impairments were expected to incur differing levels of cognitive load than sighted users, it was necessary to distinguish between those with congenital and adventitious sight loss. The visual memory of subjects in the latter category could affect the results; therefore, data from each category were analyzed separately.

Before beginning the experiment, subjects were read a description of the spatial layout of the area where they would perform the tasks and were told the nature of tasks to be performed. Subjects were given approximately 45 min for each session with a break between sessions of approximately 10 min. No special training was given, since the use of natural spoken language for input eliminated the need for expertise with any particular software. However, subjects were asked to perform a short task prior to starting the experiment to reduce effects of testing anxiety. The complexity of this task was equivalent to the simplest task in each session. No restrictions were given on the time to perform this initial task.

2.3. Experimental tasks

In each session, subjects performed a series of navigational tasks, each of which entailed planning a route, safe for pedestrians, from one location on the station to another. The program computes an initial driving route that is not optimized for pedestrians. Thus, subtasks entailed querying for conditions affecting pedestrians and modifying the route to optimize it for both length and simplicity. Data on conditions affecting pedestrians could be queried from the map database. This included road conditions such as the presence of adjacent sidewalks and crosswalks, the level of traffic and speed limits, the presence of sharp curves in the road, the amount and condition of the road shoulder, and any construction efforts underway. Also, other general conditions could be queried, e.g., weather, time of day, and locations of prominent landmarks.

Tasks were presented in series of four. Spatial complexity was increased incrementally for each task in the series to gather more data on the effects of the spatial aspect of the tasks on the results. Design of the spatial task complexity was based on techniques used in the field of Orientation and Mobility (O&M) for persons with visual impairments (Jacobson, 1993). Four basic route patterns were employed. The patterns, listed from simplest to most complex, are named by letters in the alphabet which most closely resemble their shape, i.e., "I" (straight line), "L", "U", and "Z". While these basic patterns formed the basis of the four routes, other factors, such as the number of street crossings as well as road conditions, varied the complexity within each route.

An example task using an "I" route is shown in Fig. 3. The task required the subject to determine a walking route from Headquarters to the Environmental Laboratory. An "X" indicates the beginning of the chosen path. The dashed lines indicate the direction of the final path. Double lines indicate crosswalks and sidewalks. Thus, the east side of Arkansas Road does not contain continuous sidewalks. Other relevant road conditions not visible on the diagram, but which can be queried from the database include that the road experiences heavy traffic between 7 and 8 AM, the time during which the subject is constructing the path. Note that the overall path shape resembles an "I" or straight line, i.e., the Environmental Laboratory is due south of Headquarters. However, a subject could choose two vertical street crossings and two horizontal crossings in constructing the optimal walking path to the destination.

A partial sample dialogue to construct this path is given below and illustrates some key features of user-system interactions:

System:
"For the first segment of the route, facing south toward the Environmental Laboratory, with the Headquarters Building to your left and the station entrance to your right, proceed to the east corner of Arkansas and Brazos Circle."
User:
"What is the road like from here to Environmental Lab?"
System:
"Traffic on this road is heavy from 7 to 8 AM, but there is no construction along this road currently."
User:
"Okay, um … is there a sidewalk? Will I need to cross?"
System:
"Could you repeat that please?"
User:
"Sidewalk, … is there a sidewalk … or crosswalk?"
System:
"There is sidewalk on the east side of Brazos Circle up to the corner of Brazos Circle and Arkansas."

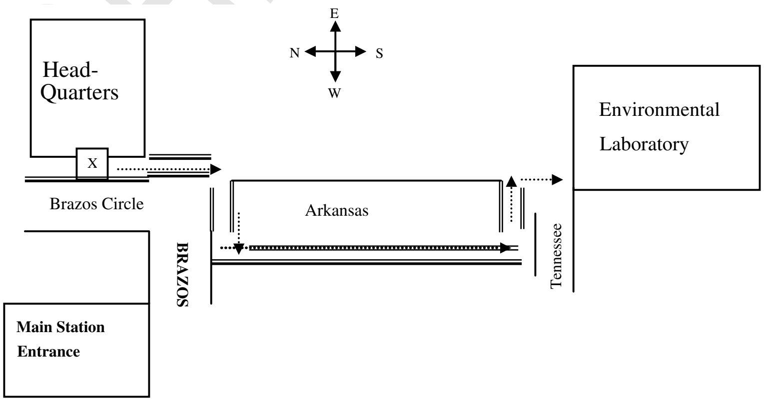

Fig. 3. Example "I" task.

User:
"How about the rest of Arkansas? Can I continue on sidewalks?"
System:
"Arkansas contains sidewalk on the west side of the street only from the corner of Arkansas and Brazos Circle to the corner of Arkansas and Tennessee."
User:
"Am I on the west side? … Where am I?"
System:
"You are at the east corner of Brazos Circle and Arkansas, facing south toward the Environmental Laboratory."
User:
"Okay cross over."

The session shown begins during the initial dialogue to establish the user's starting point and destination.

2.4. Scope of study

Issues concerning the scope and methods of the study should be mentioned, in particular those regarding the order of treatments. Presenting the visual or tactile treatment first, immediately followed by speech-only treatment, could allow the memory of the first treatment to reduce the difficulty of speech-only interaction. Ideally, the visual/tactile treatments would be presented in one session, followed by an elapsed time period of sufficient length to negate the effects of visual and tactile memory before presenting the speech-only treatments. However, time limitations required the treatments to be applied in consecutive sessions, thus, a short break of approximately 10 min was provided between each. Since this would not provide sufficient time to counter the possible effects of visual and tactile memory, the speech-only treatments were presented first. To offset possible practice effects, a warm-up session was provided. Results of this session were not analyzed. In addition, the task-level statistical tests allowed comparing results of the last task in the first session against the last task in the second session. In other words, subject performance at the time of greatest practice with the speech-only treatment could be compared against performance at the time of greatest practice with the visual or tactile treatment.

The experiments were conducted over the course of approximately three months at various academic, medical and rehabilitation agencies. Approximately 90 subjects participated in the experiments, including over 30 sighted subjects and over 60 subjects with visual impairments. As expected, a small number of experimental samples could not be analyzed. Out of the total population, data from 78 subjects were used in the analyses, including 27 sighted subjects. A variety of reasons precluded certain data from the analyses, including subjects terminating mid-session and unanticipated excessive background noise at the testing location.

3. Results

This section reviews the data analysis methodology, including the type of user and system data measured, i.e., prosodic features and recognition errors, respectively, as well as the method of measurement for each. Analyses of results are then presented comparing overall user and system data gathered in the displayless sessions to that gathered in the multimodal sessions. Next, analyses of results at the task level, i.e., comparing data from each task in displayless sessions against each task in multimodal sessions, are presented. Since spatial complexity increased with each task, results were analyzed at this level to measure the effect of the spatial complexity of the tasks on the user's prosodics, and hence cognitive load.

3.1. Data analysis

Speech data collected during the experiments was transcribed and labeled using the Tones and Break Indices (TOBI) transcription system (Silverman et al., 1992). Prosodic features were extracted and labeled per utterance by two labelers with an inter-labeler agreement of 82%. These features included: pauses (type, quantity, and length in seconds), breaths (quantity and location), fundamental frequency (F0) (maximum and minimum values), intonational phrase boundary tones (type and quantity), preboundary lengthening (in seconds), and speaking rate changes (in seconds). Acoustic data for each variable was extracted and measured per utterance. The per-utterance measurements were averaged per session as well as per task for statistical analysis. Finally, minimum and maximum F0 values per utterance were averaged per session per subject.

After the prosodic data was labeled and transcribed, matched-pair *t*-tests were performed to compare the means of the differences in the prosodic measurements in the displayless session against those measured in the multimodal session. The tests were performed comparing both overall session data as well as task-level comparisons, i.e., matched-pair *t*-tests were performed for each subject category, comparing prosodic variables for all tasks completed in displayless sessions against prosodic data for all tasks completed in multimodal sessions. Final tests were performed comparing prosodic data for the first task in the displayless session to prosodic data for the first task in the multimodal session; likewise for each subsequent task. Recognition errors and system strategies for handling them can affect the level of frustration experienced by the users and could thus impact the results. Therefore, during each session, the number and type of errors, rejection, substitution, and insertion, made by the system were measured and analyzed per utterance and then averaged per session as well as per task. Each utterance was digitally recorded and stored with an associated file containing the textual representation of the system interpretation. The digitized speech was hand-labeled orthographically during post-processing.

To reiterate, the ASR engine for the fielded system performed with an absolute WER of 10.2%. However, system understanding errors are more critical for the prototype application, since it functioned as a database query interface rather than a dictation style program. Therefore, recognition errors were analyzed on a semantic basis; hence, correct interpretation of the meaning of the user's request was considered an accurate recognition for data analysis. The reported substitution, insertion, and rejection errors are only for those utterances that resulted in an incorrect interpretation by the system. Again, system performed with an overall semantic error rate of 13.9%.

Analysis of system recognition errors on speaker utterances was conducted in a manner similar to that for the prosodic variables since identical experimental conditions were applied. Again, a matched-pair *t*-test was used to compare the means of the differences in the measurements of recognition errors extracted from the displayless session versus the multimodal session. These tests were performed to compare both overall session data as well as task-level data. In other words, matched-pair *t*-tests were performed for each subject category to compare the system recognition errors on speaker utterances for all tasks completed in the displayless sessions against those for all tasks completed in the multimodal sessions. Final tests were performed on a task-level basis, e.g., system recognition errors on speaker utterances for the first task in the displayless session were compared to those for the first task in the multimodal session; likewise for each subsequent task.

3.2. Session analyses

Several common patterns emerged in the overall session data for all categories of subjects. First, the number of hesitation pauses, i.e., those not occurring at a phrase boundary and marked "2p" in TOBI, was significantly greater during displayless sessions than multimodal sessions for all populations, at a significance level $\alpha \leqslant 0.01$. To illustrate this reduction in "2p" hesitation pauses in the multimodal session, the raw data values are plotted in Fig. 4 for one subject category, the congenitally blind, although, as stated, an equally significant reduction occurred for both the adventitious and sighted subjects. Note that while the number of "2p" pauses varies widely per individual, it is consistently reduced in the multimodal session across all subjects. In addition to the number of "2p" pauses, the average length of these pauses was significantly greater during displayless sessions than multimodal sessions for all subject categories. For sighted subjects as well as subjects with adventitious vision loss, the average length of these pauses was significantly greater during dis-

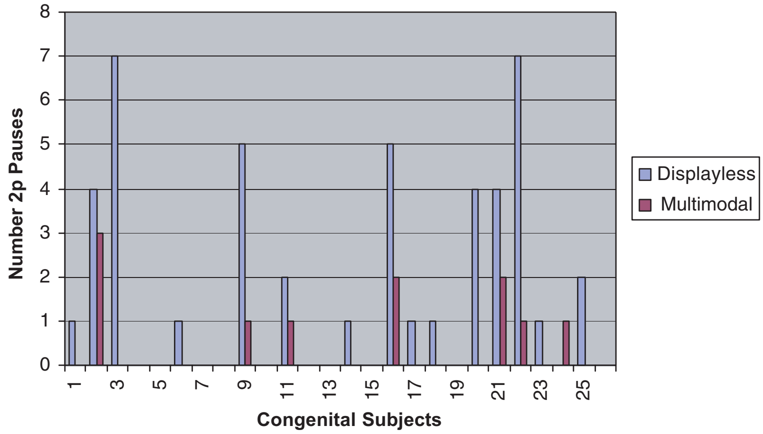

Fig. 4. Number of '2p' pauses for congenital subjects in displayless versus multimodal session.

playless sessions at the level $\alpha \leqslant 0.05$. These results indicate that this prosodic feature is not likely a good single predictor for detecting phrase boundaries.

Regarding tonal data, for all three populations, the number of low full intonational boundary tones ("L%") was significantly greater during displayless sessions at $\alpha \leqslant 0.01$. This increase presents problems for tune detection algorithms that seek to classify utterances as yes/no questions based on the ending tone in the utterance. Since significantly more utterances end in low declarative tones, it is more likely that a user may conclude yes/no questions in this manner, thus confounding algorithms expecting a high tone.

Lastly, for all three populations, the number of substitution errors made by the system on speaker utterances was significantly greater during displayless than multimodal sessions. For all other variables, results differed among subject categories. Table 1 summarizes the results, providing mean values for prosodic variables in displayless and multimodal sessions, highlighting those that differed significantly between sessions in bold with a single asterisk, "*", indicating a significance level of $\alpha \leqslant 0.05$. Table 2 provides the alpha levels for the differences in the data between sessions. A positive value represents a variable with a value that was significantly larger during the displayless session versus the multimodal session, while a negative value represents a variable with a value that was significantly smaller during the displayless session. Again, a single asterisk, "*" indicates a significance level of $\alpha \leqslant 0.05$. Note that results for subjects with congenital vision loss differ from the other two categories in certain aspects. First, the number of pauses occurring at a phrase boundary, denoted "3p", is significantly greater during displayless than multimodal sessions. Also, aspects of the tonal data differ from the other two populations. F0 values show no significant change between sessions and the number of low full intonational boundary tones, "L%", is significantly greater during displayless sessions than multimodal sessions. In addition, a larger number of durational features differ significantly between sessions. Finally, all three categories of recognition errors differ significantly between sessions for this population. Again, however, these results reflect the comparison of data from all tasks in the first session against data from all tasks completed in the second session. Task-level analyses, presented in the following section, should also be discussed.

3.3. Task-level analyses

All subjects finished at least two tasks in one or both sessions. Thus, only data from the first two tasks were analyzed at the task level. To reiterate, task-level analyses were performed to ascertain how the spatial complexity of the tasks affected the user's prosodics, and hence cognitive load.. Re-

Table 1
Mean values for all populations in overall session data analyses

	Congenital		Adventitious		Sighted	
	Displayless	Multimodal	Displayless	Multimodal	Displayless	Multimodal
Pauses						
Number 2p	***1.85****	***0.41****	***1.78****	***0.52****	***1.56****	***0.42****
Number 3p	***5.93****	***3.72****	3.61	4.22	3.84	3.31
Length 2p (s)	0.19	0.10	***0.33***	***0.16****	***0.22***	***0.07****
Fundamental freq. (F0)						
Maximum (Hz)	294	288	***316****	***261****	258	259
Minimum (Hz)	64	39	***78****	***60****	***62****	***72****
Boundary tones						
Number L%	***25****	***18****	***22****	***16****	***18****	***13****
Number H%	10	11	20	17	18	15
Durational features						
Speaking rate (words/s)	***1.6****	***1.8****	1.6	1.5	1.4	1.3
Duration (s)	3.8	3.9	3.8	3.7	***4.2****	***4.0***
Semantic error rate						
Overall	***18.4***	***14.8***	***16.9***	***10.2***	***16.7***	***11.5***
Substitution	***15.1****	***10.2****	***14.0****	***8.6****	***13.1****	***8.9****
Insertion	1.4	2.0	0.4	0.2	1.0	1.0
Rejection	2.0	1.0	2.2	1.3	2.2	1.3

* Indicates difference was significant at $\alpha \leqslant 0.05$.

Table 2
Significance of differences for all populations in overall session data analyses

	Congenital	Adventitious	Sighted
Pauses			
Number 2p	***0.0017****	***0.0089****	***0.0001****
Number 3p	***0.0256****	−0.4820	0.5428
Length 2p (s)	0.0561	***0.03260****	***0.0057****
F0			
Maximum (Hz)	0.9224	***0.0002****	0.7901
Minimum (Hz)	0.3772	***0.0492****	***−0.0040****
Boundary tones			
Number L%	***0.0001****	***0.0009****	***0.0007****
Number H%	−0.8459	0.0526	0.0584
Durational features			
Speaking rate (words/s)	***−0.0340****	0.4537	0.9971
Duration (s)	0.1206	0.3089	***0.0092****
Semantic error rate			
Substitution	***0.0163****	***0.0010****	***0.0004****
Insertion	−0.0560	0.3800	0.1249
Rejection	0.0570	0.2644	0.8591

'–' Indicates value of variable smaller during displayless session.
* Indicates difference was significant at $\alpha \leqslant 0.05$.

call that spatial complexity increases with each task; thus higher task numbers signify higher spatial complexity and greater cognitive load. Therefore, variables differing significantly for higher level tasks, e.g., Task 2, offer greater evidence that cognitive load is increased than those differing significantly for a lower level task, e.g., Task 1. Recall also that comparisons of higher-level tasks were performed to ameliorate the issue of order of treatments: subjects would have greater practice with the displayless interface once they reached the higher task levels. In other words, variables differing significantly for Task 2 provide stronger support than those found significant for Task 1 only.

Two variables differed significantly for all populations on Task 2. These included the number of hesitation pauses, denoted "2p", and the number of "L%" boundary tones, both of which were significantly greater in utterances spoken during displayless sessions than multimodal sessions. Certain patterns that characterized each population in overall session comparisons emerged in the task analyses also, but not all remained significant for Task 2. Summaries of significantly differing varia-

Table 3
Significance of differences in task-level analyses for congenital population

	Significance overall	Significance Task 1	Significance Task 2
Pauses			
Number 2p	***0.0017****	0.1364	***0.0024****
Number 3p	***0.0256****	0.3458	***0.0237****
Length 2p (s)	0.0561	0.1340	0.2915
F0			
Maximum (Hz)	0.9224	0.8828	0.6255
Minimum (Hz)	0.3772	0.9658	0.3103
Boundary tones			
Number L%	***0.0001****	***0.0319****	***0.0085****
Number H%	0.8459	0.4664	0.4038
Durational features			
Speaking rate (words/s)	***−0.0340****	***−0.0178****	***−0.6657***
Duration (s)	0.1206	0.9217	0.0861
Semantic error rate			
Substitution	***0.0163****	0.0605	0.1350
Insertion	−0.0560	***−0.0430****	0.1617
Rejection	0.0570	0.5233	***0.0250****

'−' Indicates value of variable was smaller during displayless session.
* Indicates difference was significant at $\alpha \leqslant 0.05$.

Table 4
Significance of differences in task-level analyses for adventitious population

	Significance overall	Significance Task 1	Significance Task 2
Pauses			
Number 2p	***0.0089****	0.2326	***0.0138****
Length 2p (s)	***0.03260****	0.4727	***0.0285****
F0			
Maximum (Hz)	***0.0002****	***0.0206****	***0.0081****
Minimum (Hz)	***0.0492****	***0.0428****	0.9680
Boundary tones			
Number L%	***0.0009****	***0.0009****	***0.0189****
Number H%	0.0526	***0.0526****	0.2285
Durational features			
Speaking rate (words/s)	0.4537	0.1892	0.4819
Duration (s)	0.3089	0.2070	0.9189
Semantic error rate			
Substitution	***0.0010****	***0.0178****	***0.0015****
Insertion	0.3800	0.6639	0.0881
Rejection	0.2644	0.1777	0.3819

'−' Indicates value of variable was smaller during displayless session.
* Indicates difference was significant at $\alpha \leqslant 0.05$.

bles at the task level for all populations are given in Tables 3–5.

For subjects with congenital vision loss, an increase in the average length of hesitation pauses, denoted "2p", occurring in utterances from displayless versus multimodal sessions was not found significant for either Task1 or Task 2. However, the number of "3p" pauses, occurring at a phrase boundary, was significantly greater in utterances from displayless sessions than multimodal sessions for Task 2 only. Speaking rate as well as duration of utterance did not differ significantly for Task 2. Although all categories of recognition errors differed significantly in overall session comparisons, only rejection errors were significantly greater for Task 2 during displayless sessions. The significant differences between sessions per task for this population are summarized in Table 3.

For subjects with adventitious vision loss, maximum F0 was significantly higher in utterances for Task 2 during displayless sessions than multimodal sessions. Results for this population are summarized in Table 4. The minimum F0 was significantly higher for Task 1 only. The number of "H%" boundary tones did not remain significantly higher for Task 2 during displayless versus multimodal sessions, although it was significant for Task 1. The number of high intermediate boundary tones, denoted "H-", was significantly greater for Task 2, although this variable did not differ in overall comparisons. The number of substitution errors occurring for utterances in displayless rather than multimodal sessions was significantly greater for Task 1 and Task 2.

Results for sighted subjects are given in Table 5. In contrast to the adventitious population, minimum F0 was significantly lower in utterances for Task 2 during displayless sessions, but maximum F0 did not differ significantly between sessions. Other tonal changes include the number of "H%" boundary tones, which was significantly greater in utterances for Task 2 from displayless sessions. Finally, the number of substitution errors

Table 5
Significance of differences in task-level analyses for sighted population

	Significance overall	Significance Task 1	Significance Task 2
Pauses			
Number 2p	***0.0001****	***0.0233****	***0.0013****
Length 2p	0.0057	0.1034	***0.0021****
F0			
Minimum (Hz)	***–0.0040****	***–0.0061****	***–0.0057****
Maximum (Hz)	0.7901	0.7536	0.8606
Boundary tones			
Number L%	***0.0007****	***0.0209****	***0.0006****
Number H%	0.0584	0.9889	***0.0450****
Durational features			
Duration (s)	***0.0092****	0.0750	***0.0050****
Speaking rate (words/s)	0.9971	0.1860	0.4381
Semantic error rate			
Substitution	***0.0004****	0.1307	***0.0072****
Insertion	0.1249	0.2352	1.0000
Rejection	0.8591	1.0000	0.1675

'–' Indicates value of variable was smaller during displayless session.
* Indicates difference was significant at $\alpha \leqslant 0.05$.

was significantly greater for Task 2 only during displayless versus multimodal sessions, at the significance level $\alpha \leqslant 0.01$.

4. Discussion of results

One conclusion that can be drawn from the analysis is that hesitation pauses are increased, for all categories of users, in the displayless condition. This indicates a likely increase in the amount of cognitive effort and planning required to use the displayless navigational interface. This additional effort must be counterbalanced for widespread acceptance of these interfaces to occur. Further, the increase in hesitation pauses appears to have increased the number of misrecognition errors made by the system, which in turn negatively affects the level of user satisfaction with the interface.

The dissimilarities in the results for the congenital population from those of the sighted and adventitious population provide insight regarding the relationship between prosodics and recognition error rate. The congenital population exhibited fewest differences in tonal variables, i.e., F0 values and intonational boundary tones, between sessions. In addition, for this population only, substitution errors did not significantly increase during displayless sessions. Conversely, the latter two populations exhibited the largest number of differences in tonal data between sessions, significant increases in the length of hesitation pauses, as well as a significant increase in substitution errors during displayless sessions. These results suggest that the combination of intonational changes and hesitation pauses most significantly affected the substitution error rate. No correlation between disfluencies and recognition error rate was found in a study conducted by Rosenfeld et al. (1996). However, the study measured disfluencies, not pauses exclusively. In addition, the application entailed the predominant use of monosyllabic phrases, rather than the natural language queries used in this research. The differences in the application as well as the prosodic variables measured increases the value of a study using data from this research to examine the relationship between prosodics and recognition error rate.

All populations analyzed in this research exhibited significant differences for at least one prosodic feature when using the displayless interface; for sighted and adventitious populations, a combination of prosodic features differed significantly. These results support the use of multiple features for robust prosodic pattern detection for displayless navigational applications. In particular, the universality of results concerning pauses provides evidence that this prosodic feature is not likely a good single predictor for phrase boundaries. The differences in tonal and durational data, particularly for the sighted and adventitious populations, indicate that these features are also important for phrase boundary detection algorithms.

Further, the differences in boundary tones, particularly the significant increase in "L%" tones during displayless sessions, present problems for tune detection algorithms which seek to classify utterances as yes/no questions based on the ending tone in the utterance. Since significantly more utterances end in low declarative tones, it is more likely that a user may conclude yes/no questions in this manner, thus confounding algorithms expecting a high tone. Finally, similar problems arise for prominence detection algorithms that rely on a single acoustic cue, such as F0, to detect the speaker's emphasis. Given the variability in prosodic features during displayless sessions, a speaker may more likely use a combination of cues to indicate emphasis during these sessions, such as durational lengthening along with shifts in F0.

Much of the work in prosodic pattern detection has relied on the use of either recorded speech read from a prepared text or from interactions with a speech surrogate. This work adds to the limited number of studies that were conducted using these conditions. Only recently have studies using spontaneous speech with a live recognizer, such as the DARPA EARS (2003) program, been reported, and findings of these studies are not yet conclusive.

5. Conclusions and future work

This research examined the assumption that the prosodics of user speech produced in sessions employing a displayless interface would differ significantly than that produced employing a multimodal interface. For all categories of subjects, significant differences in certain prosodic features were found, including hesitation pauses and low L% boundary tones. Further, for sighted and adventitious populations, the combination of tonal differences and increased hesitation pauses appears correlated to the increased substitution error rate for these users.

This study used significant variations in prosodics during displayless sessions to measure increases in cognitive load. Thus, each population experienced some additional cognitive load without a visual or tactile display since each exhibited significant variations in certain prosodic variables during displayless sessions. However, subjects in the sighted and adventitious populations experienced the most additional cognitive load when using a speech-only interface since they exhibited the most prosodic variations during displayless sessions. Conversely, subjects in the congenital population experienced the least additional cognitive load when using a speech-only interface, since they exhibited the least prosodic variations during displayless sessions. This could possibly be attributed to a lack of visual memory and thus, a lack of frustrated attempts to "visualize" the geographical area while problem solving. However, since such a hypothesis was not formally investigated in this research, further study of the issue is needed to confirm or disprove it.

Regardless of the cause in dissimilarities, decreasing cognitive load for all populations of displayless interface users is important. Difficulty in simply maintaining a general sense of compass directions appeared to contribute greatly to the increase in cognitive load during displayless sessions. The prototype program provides explicit compass directions in relation to the user's current position as well as whether to turn left or right, or continue. Nonetheless, subjects could be observed repeatedly "interpreting" these instructions with respect to their current location. Many subjects demonstrated through a variety of physical mannerisms, including verbalizing, e.g., "If south is to my left," gesturing, e.g., outlining a position in the air with the fingers, or for sighted subjects, closing eyes to "visualize" the area in question. Some methods to reduce such cognitive effort include the integration of palm-size or head-mount displays, where possible, or the use of non-speech audio cues. For the latter, stereo localization cues conveying the direction of travel showed promise in research described by Loomis et al. (1994).

The results of this research also provide evidence that single acoustic cues are not robust predictors in prosodic pattern detection. These issues can be explored further from the database of spontaneous speech produced by the investigation. Particular questions of interest to evaluate include the use of pauses in phrase boundary detection, the use of F0 for emphasis, and the use of high versus low declarative tones for posing yes/no questions.

Lastly, the results revealed potential human factors problems, i.e., increases in cognitive load, which must be addressed to ensure the success of displayless navigational interfaces. In addition, this study gathered baseline observations of the variables that contributed to the increase in cognitive load. These observations can serve as a foundation for improving the usability of these interfaces. The most salient observation pertained to users' difficulty in maintaining a general sense of compass directions. Solutions to explore include augmenting the interface with localized sound sources and/or a palm-sized visual or tactile map.

A final area for future investigation pertains to the nature of the prototype deployment. The experiment described in this research deployed the prototype in a stationary mode in an office environment. Deployment in a mobile environment with the noise and distractions of a live situation could yield different results. This study attempted to isolate the spatial and verbal aspects of the navigational problem. However, the results of this study compared to those from a study conducted in a mobile environment could provide a richer knowledge source than either alone.

In conclusion, displayless navigational technology offers many potential benefits to the user community. Perhaps of greatest value, it offers the possibility of a higher degree of independence in daily activities to all users, whether constrained by the environment or visual acuity. This research examined and illuminated many issues critical to the successful delivery of this technology.

Acknowledgment

Special thanks are due to the rehabilitation agencies that allowed testing for this research, including the Rehabilitation and Training Center for Blindness and Low Vision at Mississippi State University, the Addie McBryde Rehabilitation Center for the Blind in Jackson, MS, Lion's World in Little Rock, AR, and the Louisiana Center for the Blind in Ruston, LA.

References

Baca, J., 1998. Displayless access to spatial data: Effects on speaker prosodics. Doctoral dissertation, Mississippi State University, published as WES Technical Report ITL-98-3.

Baca, J., Zheng, F., Gao, H., Picone, J., 2003. Dialog systems for automotive environments. In: Proc. Eurospeech, Geneva, Switzerland.

Barth, J., 1983. Tactile Graphics Guidebook. American Printing House for the Blind, Louisville, Kentucky.

Bayer, S., Bernstein, E., Duff, D., Hirschman, L., LuperFoy, S., Peet, M., 1995. Spoken language understanding report on the Mitre Spoken Language System. In: Proc: Spoken Language Systems Technology Workshop, January 22–25, 1995, pp. 243–251.

Buhler, D., Minker, W., Haubler, J., Kruger, S., 2002. Flexible multimodal human–machine interaction in mobile environments. In: Proc. ICSLP '02, Denver, CO, USA.

Campbell, W.N., 1992. Prosodic encoding of speech. In: Proc. ICSLP'92, Banff, Canada, pp. 663–666.

Chen, F., Withgott, M., 1992. The use of emphasis to automatically summarize a spoken discourse. Proc. Internat. Conf. on Acoust., Speech Signal Process. (ICASSP), Vol. 1. IEEE, New York, pp. 229–232.

Dahl, D., Bates, M., Brown, M., Fisher, W., Hunicke-Smith, K., Pallett, D., Pao, C., Rudnicky, A., Shriberg, A., 1994. Expanding the scope of the ATIS task: the ATIS-3 corpus. In: Proc. Human Language Technology Workshop, March 8–11, 1994, Plainsboro, NJ, pp. 43–48.

Daly, N., Zue, V., 1990. Acoustic, perceptual, and linguistic analyses of intonation contours in human/machine dialogues. In: Proc. ICSLP'90, Kobe, Japan, pp. 497–500.

DARPA, 2003. DARPA EARS Conference, Boston, MA 21–22, 2003.

Godfrey, C.H. J., Doddington, G., 1990. The ATIS Spoken Language Systems corpus. In: Proc: Speech and Natural Language Workshop. Morgan Kaufman, Hidden Valley, PA, pp. 96–101.

Huber, D., 1989. A statistical approach to the segmentation and broad classification of continuous speech into phrase-sized information units. In: Proc. Internat. Conf. on Acoust., Speech, Signal Process. (ICASSP). IEEE, Glasgow, Scotland, pp. 600–603.

Jacobson, W.H., 1993. Basic outdoor O&M skills. In: The Art and Science of Teaching Orientation and Mobility to Persons with Visual Impairments. AFB Press, New York, NY, pp. 105–116.

Kamm, C., 1994. User interfaces for voice applications. In: Voice Communication Between Humans and Machines. National Academy Press, Washington, DC.

Kozlowski, L., Bryant, K., 1977. Sense of direction, spatial orientation, and cognitive maps. J. Experiment. Psychol. 3 (2), 590–598.

Loomis, J.M., Golledge, R.G., Klatzky, R.L., Speigle, J., Tietz., J., 1994. Personal guidance system for the visually impaired. In: Proc. ASSETS 94, ACM Conference on Assistive Technologies, Los Angeles, CA, pp. 85–91.

Nakai, M., Shimodaira, H., Sagayma, S., 1994. Prosodic phrase segmentation based on pitch-pattern clustering. Electron. Comm. Jpn 77 (6), 80–91.

Noth, E., Batliner, Kieblingm, A., Kompe, R., 2000. VERBMOBIL: the use of prosody in the linguistic components of a speech understanding system. IEEE Trans. Speech Audio Process. 8 (5), 519–531.

Okawa, S., Endo, T., Kobayashi, T., Shirai, K., 1993. Phrase recognition in conversational speech using prosodic and phonemic information. IEICE Trans. Inform. Syst. E76-D (1), 44–50.

Paul, D., Baker, J., 1992. The design of the Wall Street Journal-based CSR Corpus. In: Proc. ICSLP '92, Banff, Alberta, Canada, pp. 899–902.

Pellom, B., Ward, W., Hansen, J., Hacioglu, K., Zhang, J., Yu, X., Pradhan, S., 2001. University of Colorado Dialog Systems for Travel and Navigation, In: Proc. of the 2001 Human Language Technology Conference (HLT-2001), San Diego, CA.

Rosenfeld, R., Byrne, B., Iyer, R., Liberman, M., Shriberg, L., Unveferth, J., Vidal, E., Agarwal, R., Vergyri, D., 1996. Error analysis and disfluency modeling in the Switchboard domain. In: Proc. ICSLP'96, Philadelphia, PA, SAP1S1.3.

Scherer, K.R., 1981. Speech and emotional states. In: Speech Evaluation in Psychiatry. Grune-Stratton, New York, pp. 189–220.

Scherer, K.R., Grandjean, D., Johnstone, T., Klasmeyer, G., Banziger, T., 2002. Acoustic correlates of task load and stress. In: Proc. ICSLP'02, Denver, CO, USA.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., Hirschberg, J., 1992. TOBI: a standard for labelling English prosody. In: Proc. ICSLP'92, Banff, Alberta, Canada, pp. 867–870.

Thorndyke, P., Stasz, C., 1980. Individual differences in procedures for knowledge acquisition from maps. Cognitive Psychol. 12, 137–175.

Waibel, A., 1988. Prosody and Speech Recognition. Morgan Kaufmann, San Mateo, CA.

Wightman, C.W., Ostendorf, M., 1994. Automatic labeling of prosodic patterns. IEEE Trans. Speech Audio Process. 2 (4), 469–481.

Woodland, P.C., Odell, J.J., Valtchev, V., Young, S.J., 1994. Large vocabulary continuous speech recognition using HTK. In: Proc. ICASSP '94, Phoenix, AZ, USA, pp. II/125–II/128.