Journal of International Technology and Information Management

Volume 17
Issue 3 *Double Issue 3/4*

Article 7

2008

uC: Ubiquitous Collaboration Platform for Multimodal Team Interaction Support

Veton Z. Kepuska
Florida Institute of Technology

Sabri Gurbuz
Cognitive Information Science Laboratories Kyoto

Walter Rodriguez
Florida Gulf Coast University

Stephen M. Fiore
University of Central Florida

Deborah Carstens
Florida Institute of Technology

See next page for additional authors

uC: Ubiquitous Collaboration Platform for Multimodal Team Interaction Support

Authors

Veton Z. Kepuska, Sabri Gurbuz, Walter Rodriguez, Stephen M. Fiore, Deborah Carstens, Patrick D. Converse, and David Metcalf

uC: Ubiquitous Collaboration Platform for Multimodal Team Interaction Support

Veton Z. Këpuska
Florida Institute of Technology
USA

Sabri Gurbuz
NICT/ATR, Cognitive Information Science Laboratories Kyoto
JAPAN

Walter Rodriguez
Florida Gulf Coast University
USA

Stephen M. Fiore
University of Central Florida
USA

Deborah Carstens
Patrick D. Converse
Florida Institute of Technology
USA

David Metcalf
University of Central Florida
USA

ABSTRACT

A human-centered computing platform that improves teamwork and transforms the "human-computer interaction experience" for distributed teams is presented. This Ubiquitous Collaboration, or uC ("you see"), platform's objective is to transform distributed teamwork (i.e., work occurring when teams of workers and learners are geographically dispersed and often interacting at different times). It achieves this goal through a multimodal team interaction interface realized through a reconfigurable open architecture. The approach taken is to integrate: (1) an intuitive speech- and video-centric multi-modal interface to augment more conventional methods (e.g., mouse, stylus and touch), (2) an open and reconfigurable architecture supporting information gathering, and (3) a machine intelligent approach to analysis and management of heterogeneous live and stored sensor data to support collaboration. The system will transform how teams of people interact with computers by drawing on both the virtual and physical environment.

INTRODUCTION

"Collaboration software now provides the greatest value for workers."
Wall Street Journal, September 18, 2007

The key to the success of any complex system, such as the one described in this paper, is development of the multimodal user interface (MMUI) that enables intuitive and easy access to and use of systems features. One of the most natural interface that is currently underutilized is human speech. However, accurately predicting users' intent from speech, prosodic and visible speech features, and other features (e.g., touch-screen and/or mouse/stylus movement) is largely an unsolved problem. The ability to solve this problem rests on the already developed Wake-Up-Word (WUW) speech recognition (SR) system. In this system, we will integrate a model of higher level cognitive decision-making to discriminate between an ***alerting*** context (e.g., "**Computer** - show me the chart!) and a ***referential*** context (e.g., "My **computer** has a quad-core Intel processor"). This MMUI model will serve as a basis for the proposed collaborative Ubiquitous Colaboration or uC "you see" platform. In addition, an understanding of ubiquitous collaboration, particularly one that effectively integrates tools and methods arising out of engineering, modeling and simulation, and the organizational and behavioral sciences is will be developed. Ultimately, the goal is to incorporate design features that will ensure that the uC will enhance collaboration in work environments and team performance by optimizing the ***natural interaction*** between the users and the system.

Successful integration of WUW SR technology in MMUI will impact not only the proposed uC platform but also enable development and deployment of this SR technology in numerous other applications not possible today (e.g., smart court rooms, smart conference rooms and smart homes) as well as enable people with disabilities, and more generally, change the way humans interact with their computerized systems. While commercial SR technology has been around for over 20 years, with augmented WUW technology human-machine interaction will be revolutionized. This human-centered interfacing will lead one step forward in providing the means for applications that interact with users in a natural manner.

The uC system will transform distributed teamwork in that (1) it will acquire real-time data from embedded sensors and feed that data into the appropriate work/learning opportunity and/or problem solving challenge; (2) it will do this in the context of team-task roles and their interdependencies (i.e., sensors help populate and make more meaningful the context in which the team is interacting whether it be 3D, 2D, voice, its transcription, or action triggered by it); (3) the sensors augment reality such that real-time information (e.g., storm surge height/locations) is inextricably linked with the visual simulation - the simulation is not based upon pre-canned models; (4) while some existing systems help people collaborate online, few of them have functionality that infuse real-time environmental "knowledge" into that interaction in a way consistent with the principles of team/task work; and (5) none use context-based speech recognition and visualizations.

The first part of this paper presents a case for the uC platform. Subsequent sections explain the project scope, methodology, tasks and collaboration plan.

A CASE FOR UBIQUITOUS COLLABORATION PLATFORM FOR MULTIMODAL INTERFACE TEAM INTERACTION SUPPORT

The workplace and edu-space of the future is rapidly evolving into distributed workgroups. The members of the workgroups are becoming team-learners that must overcome the barriers created by geographical distance and/or time in order to effectively perform their tasks. Going beyond the present functionality of human-centered computational systems, an integrated generic ubiquitous collaboration platform of the future must incorporate the following features: (1) *a multi-modal user interface affording intuitive and easy access to and execution of even the most complex features of the platform;* (2) *an open and reconfigurable hierarchical architecture;* (3) *access to all live and/or stored distributed data sources concurrently or individually anytime anywhere (ubiquitous);* (4) *facilitation of synchronous and/or asynchronous multimodal information sharing by utilizing the visual configuration manager and sharing the settings among team members;* (5) *support for improved management and efficient decision making via machine learning features for forecasting and decision support through multimodal/multisensory data mining and analysis;* (6) *facilitation of enhanced collaboration between co-located and/or geographically distributed team members;* and (7) *facilitation of total knowledge management.*

To achieve the goal of creating a multimodal team interaction uC platform that will transform the human-machine interaction experience the following tasks must be researched, developed, integrated and addressed: (1) *a speech- and video-centric multi-modal interface to augment more conventional methods (mouse, stylus and touch),* (2) *an open and reconfigurable architecture supporting information gathering, and* (3) *an approach to analysis and management of heterogeneous live and stored sensor data to support decision making*.

By performing the above tasks, the emergence of an important new facet of this team performance and knowledge management technology through development of a platform that supports and enhances collaboration is expected. Succinctly, we will:

- Explore and Develop an Intuitive Multimodal User Interface (MMUI) – This MMUI will use a computational modeling approach that infers information and intent from speech, acoustic prosodic and visible speech and other features (e.g., gesture, posture, head motion, eye movement, and other more conventional interfaces such as touch-screen, mouse, and stylus driven interfaces) for a higher level cognitive decision which discriminates between an ***alerting*** context and a ***referential*** context thus modeling intelligence in mimicking robustness of human speech perception and intention understanding. The "Wake-Up-Word" (WUW), Këpuska (2006), Speech Recognition System (Këpuska & Klein, 2008) and Klein (2007) will be used as a basis for this research and development effort. This interface will enable intuitive and natural interaction between users and the system using only voice or combined multimodal interfaces when appropriate to carry out complex simultaneous tasks;

- Explore and Develop the Integration of Visual Simulation – This is a web-based feature that will be based on technologies already developed by one of the authors, (Rodriguez, Opdenbosch & Satamaria, 2006). Additional research will be undertaken to explore the correlated interoperation and expansion of features targeting a variety

of computing platforms, including mobile computing and telephony devices as an example of human centered computing. The uCollaborator.org coalition partners have developed powerful data acquisition and database connectivity features that will be integrated in the uC platform;

- Explore and Develop Distributed Network Sensors – These capabilities would provide embedded and networked sensors to improve team process and performance by helping to reduce the risks, uncertainty and variability that often arises when real-time data is not provided to all team-members in a timely fashion;

- Explore and Develop Team Performance Improvement Capabilities – These process scaffolding capabilities would provide portable tools to augment dynamic problem solving and improve performance. Specifically, these features envision a web-based environment that integrates distributed sensors with team interaction scaffolds to support process and performance; and

- Explore and Develop Usability Evaluation – The evaluation would ensure usability to enhance team performance in heterogeneous socio-technical systems, such as the proposed uC platform.

In sum, distributed teamwork continues to impact competitiveness as collaboration across time and space becomes increasingly frequent in the workplace and the edu-space. Yet, system technology has not been fully developed that integrates the potential for ubiquitous team performance and multimodal team interaction support. While there are groupware systems on the market, coordinating teams in an efficient and effective manner continues to be a major challenge for both business and academia. Currently collaboration is supported piecemeal by groupware and courseware designed to support communication and coordination activities among distributed co-workers or distance-learners. However, none have leveraged the potential utility of distributed and embedded visual-simulation, speech recognition, and sensor technologies for supporting real-time collaboration. Thus, ***a unified ubiquitous collaboration system still remains to emerge, particularly one that effectively integrates tools and methods arising out of engineering, modeling and simulation, and the organizational and behavioral sciences***. By integrating best practices from the science of teamwork with the latest in visual-simulation, speech-recognition and distributed sensor technologies, uC will revolutionize distributed team performance. *Beyond groupware features, uC will enhance collaboration and team performance via the integration of visual simulation, networked-sensors, distributed data acquisition and speech and multi-modal recognition.*

INTEGRATED UC PLATFORM AND ITS FEATURES

Vision: uC technologies will be developed to connect the physical and virtual worlds by gathering real-time data and, more importantly, connect the specialized knowledge of team members, even if they are separated by time and space. The system platform and associated devices will connect co-located teams of people with individuals dispersed throughout various geographic locations. Succinctly, uC will transform the traditional workplace into an efficient and effective team-space. This research effort will explore geographical and temporal fragmentation as well as data collection, data distribution and data visualization via remote

networked sensors, visual simulations, voice recognition, among many other functions. For instance, while accessing and receiving real-time data embedded in the physical world, users would be able to see all the team members as they chat synchronously (same time) or query each other asynchronously (different time) working together to solve a complex problem and arriving at a collective decision.

Physical and Logical Design

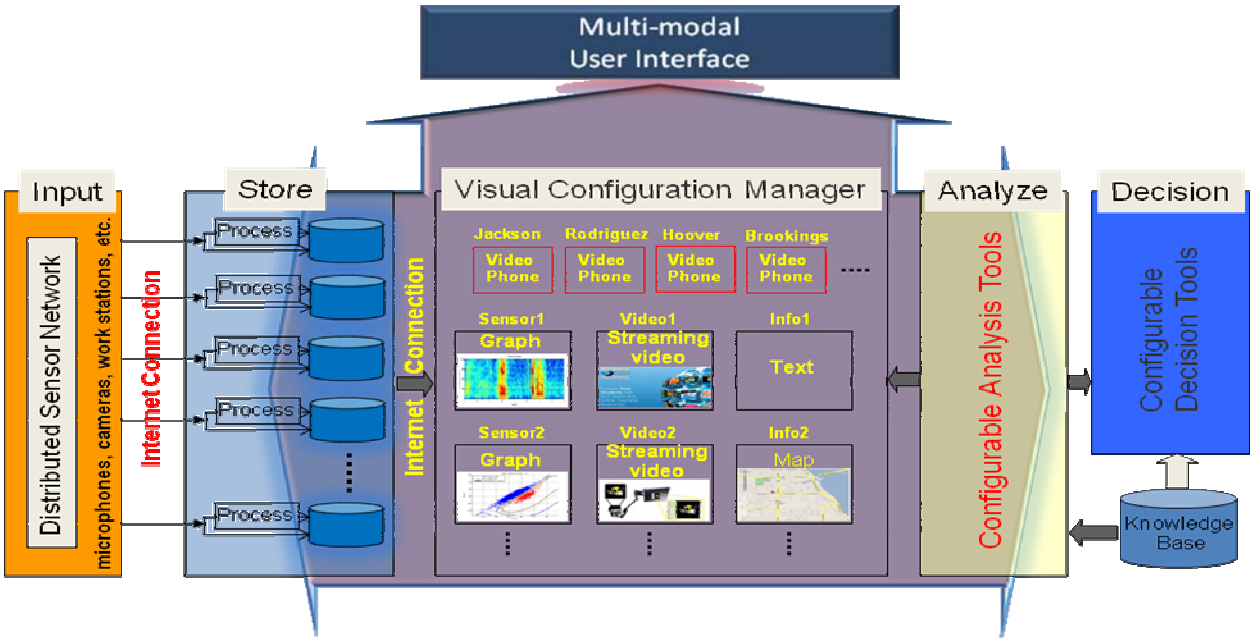
As an integrated hardware-software-network open platform, uC devices and systems will be designed in different sizes and configurations. For instance, a basic conference room tabletop unit may be visualized as a flattened cylindrical shape where the sides are standard LCD panels or foldable (collapsible) LCD segments (similar to i-Phones connected or hinged together side-by-side, so they can pivot at an angle.) The uC processing unit may be located in the middle with a protruding telescopic post holding a 360 Immersive Media camera to capture the image of team members (real-time image). For users in a different time zone, the system will provide asynchronous mode capabilities by storing prerecorded video or avatar simulations. Users may be able to select the way that video images and workspace visualization data are displayed on the LCD screens. For example, teams at various sites may be selected to appear on rings (bands) of the display so that anyone seated on opposite sides of the table will be able to see the participants at various locations. This tabletop unit may be located in the center of the conference table (at each locale) with engineers around the table or against a wall. The tabletop system may have a handle for easy transportation, while the smaller portable devices look similar to laptops or interconnected (foldable) PDAs with a built-in telescopic camera. However, initially the uC prototype system will be designed for Windows/Intel based PC's and standard web cameras.

Architecture

The uC system will provide a ubiquitous (anytime, anywhere) environment realized through mobile and fixed technologies and scaffolded by group support software. The core of this integrated platform is a collaboration engine consisting of an architecture that supports both generic collaborative processes along with task specific team processes instantiated through a sophisticated suite of advanced modular technologies. This uC engine drives dynamic and real-time collaborative problem-solving and decision-making by integrating sensor and human data from the physical world on location in the field with group support software (groupware) that efficiently and effectively manages team interaction. The system will be designed using rapid-prototyping and concurrent design methodologies so the product and system processes are built simultaneously.

The proposed open architecture has five reconfigurable layers: input layer, process and storage layer, visualization layer, analysis layer, and decision layer, as depicted in Figure 1.

Figure 1: Software Architecture of WUW-driven general Human-Machine Interactive System.



The input layer can be configured with one or more distributed networked sensors, cameras, work data and yield data streams. The process and storage layer processes its inputs (e.g., speech signals are transcribed, and video streams are analyzed) and stores the raw data and processed data. The visualization layer allows the user to configure the display window by drag-and-drop or pull-down menu selections. The analysis layer analyzes the data with configurable analysis tools and techniques. The decision layer suggests a decision or work plans based on its current configuration. All the configurations are decided by the user for an application. An application can be for scientific research project coordination (i.e., coordinating projects such as this one), solar photovoltaic power plant design and construction (i.e., http://fgcusolar.com/home), hurricane and emergency communications, healthcare management, financial or physical infrastructure monitoring, and many other team-based decision-making and complex operations.

This uC project is part of the uCollaborator.org coalition efforts pursuing a systematic program of research and development in order to seamlessly integrate extant theory on team process and performance with developing technologies. The organizations conceptual technology development plan in relation to process and performance issues associated with collaboration is presented in Table 1. This illustrates the need for a holistic approach to the solution by recognizing that technologies, processes and content all need to be taken into account for distributed collaboration to succeed.

Table 1: uC Team-Process Components.

Concepts	Technology	Process	Content
	Voice & Visual Voice Recognition	Communication and Coordination	Data from Sensors
		Usability Evaluation	Graphic/Visualizations on Simulation
	Visual Simulation		
	Integrated Sensors	Decision Making and Problem Solving	Text/Images supporting any of the processes
	Mobile/Hand-held		

While some of these technologies exist (e.g., Cisco's TelePresence system [Cisco TelePresence], HP Halo telepresence and videoconferencing solution [HP Halo]), they are either underutilized or used in a piece-meal fashion, and a significant amount of research and development needs to be done to fully integrate them to create truly distributed collaborative technologies. A more detailed discussion of uC characteristics, features, and capabilities to be explored, researched, and developed is provided next.

UC METHODOLOGY, FRAMEWORK AND TEST-BED SYSTEM

The collaborative team performance methodology, framework and test-bed system will be designed to support real-world complex problem-solving. The uC system will involve a family of new devices and integrated systems for: (1) intuitive, simple and natural multi-modal (voice, video, touch and mouse) user interfaces facilitate ease of use of all system features which otherwise would be difficult to master and use effectively, (2) sensing, collaborating, analyzing and responding to rapidly changing requirements and demands, and (3) making real-time decisions under risk and uncertainty. uC tools are based on visualizing quantitative and qualitative data (via integrated dashboards). It is the authors' strong belief that such a system can only be of practical value if it will be easy to use. Therefore, key to its success, due to the necessity of supporting inherently complex and context-dependent interactions among the users and the system, is in the development of an effective multi-modal user interface.

There are several activities that will be undertaken to make uC a viable test-bed system. We are planning to use "Agile Methodology" and "Rapid Application Development (RAD)" methodologies to develop the test-beds and prototypes. Use of RAD, in this domain, will provide insights to companies that desire to improve software productivity.

In the past decade, a radically novel proprietary Speech Recognition system that surpasses HTK (the leading academic SR system) from 1449.75% to 15166.15%, and Microsoft's SDK5.1 (the leading commercial SR system) from 919.67% to 1760.00% measured as overall relative error rate improvement was developed, Këpuska and Klein (2008). This performance accuracy will enable the development of intuitive voice interfaces that otherwise would not be possible, Këpuska, (2007). Specifically, this technology enables robust statistical methods capable of recognizing with high accuracy (Correct Acceptance > 99%) a specific word/phrase, called Wake-Up-Word (WUW) SR, with practically no false alarms (False Acceptance ≈ 0%), Këpuska

and Klein (2008), and Klein (2007). In addition, an HMM based audio-visual speech recognizer will be integrated. Within this framework, the goal is to advance human computer interactions for extending the convenience and utilization of computing for new applications such as the one in this proposed project.

The uC system will utilize the 3-D nViewer system that XYZ Solutions Inc markets. Succinctly, Visual Simulation is a web-based uC feature that has been fully developed with NSF funding and further implemented by partner companies. Theoretical issues related to team process and performance, Fiore and Salas (2007); Salas and Fiore (2004), will be integrated into the uC platform to facilitate problem solving in a variety of contexts. New research is being undertaken to explore correlated interoperation and expansion of features on a variety of platforms, including mobile computing and telephony devices. The system will adopt the most recent developments in technology-supported learning with mobile technologies (Metcalf, 2006, 2002; Lea, Yi & Kannan,2007).

The uC client-server architecture is envisioned as a research platform to investigate and optimize several visualization questions principally brought about due to multiple users employing heterogeneous hardware and software platforms. Resolving these issues will ensure that the uC team performance system would effectively manage and support the type of complex problem-solving experienced in modern engineering. Specific areas of research are described in the following sections.

INTEGRATION OF AUTOMATIC SPEECH RECOGNITION SYSTEM INTO UC PROTOTYPE

Goal

This task entails development of a friendly and intuitive uC context-dependent interface via Automatic Speech Recognition (ASR). ASR enables a computer to convert a speech audio signal into its textual transcription, Huang et al. (2001). While many tasks are better solved with visual, pointing interfaces, or keyboards, speech has the potential to be a better interface for a number of tasks where full natural language communication is useful, Zue and Cole (1997) and the recognition performance of the SR system is sufficient to perform the tasks accurately, Këpuska (2006, 2007, 2008), Klein (2007 a, b, c). This includes hands-busy or eyes-busy applications, such as where the user has objects to manipulate or equipment/devices to control as envisioned usages of uC. Some motivations for building ASR systems are: to improve human-computer interaction through spoken language interfaces, to solve difficult problems such as speech to speech translation, and to build intelligent systems that can process spoken language as proficiently as humans.

Speech as a computer interface has numerous benefits over traditional interfaces such as a GUI with mouse and keyboard: speech is natural and intuitive for humans, requires no special training, improves multitasking by leaving the hands and eyes free, and is often faster and more efficient than using conventional input methods. These features of uC will make the collaboration process more natural and will support the problem solving processes engaged by the team-learners that otherwise would not be possible to achieve.

Expanding the capabilities of the current WUW SR system and integrating it within uC to provide a natural interface is one of the major components of this research. Features of the current WUW SR will be augmented with all prosodic features such as intonation, rhythm, tempo, loudness, and pauses to enhance and make more robust identification and recognition of ***WUW context***. All prosodic features can be generated by only three parameters: fundamental frequency (F_0), energy, and duration (Milone &Rubio,2003). Based on our personal experience and other research findings, e.g., Pinker (1994), discovering and quantifying a clear paradigm describing the ***attention grabbing word/phrase or alerting*** context (i.e., WUW context) necessary for intuitive and natural human-machine interaction is expected. This finding will be used to develop specific prosodic models that would augment our WUW SR system to discriminate with high reliability between a target word/phrase having a ***wake-up*** or ***alerting*** context (e.g., "**Computer**, begin Power-Point Presentation.") from that same word/phrase in a ***referential*** context (e.g., "My **computer** has a quad-core processor."). The augmented system will be more flexible, enabling a more natural and intuitive interaction with the computing system while preserving the robustness and reliability of the original SR system.

HUMANS, IN ADDITION TO USING PROSODIC CLUES IN LANGUAGE COMPREHENSION, USE ADDITIONAL VISUAL SENSORY INPUT TO DETERMINE THE CONTEXT OF DIALOG. THUS, WE WILL ALSO STUDY VIDEO INPUT (SEE NEXT SECTION

Aiding Determination of Alerting Context with Integration of Visual Clues) to discover visual clues that pertain to the wake-up context. Fusing audio and video features is expected to add further robustness and flexibility to the system as well as capabilities that add to the naturalness of the user interface. Initially we will conduct these studies using existing movie clips. However, given evidence from past studies indicating that human behavior may change when it involves interacting with a computer (in comparison to interacting with other humans; e.g., see Dahlbäck, Jönsson, and Ahrenberg (1998), we will also use a Wizard-of-Oz approach in developing and assessing these uC capabilities (discussed in the Evaluation of uC-User Interaction section).

Approach

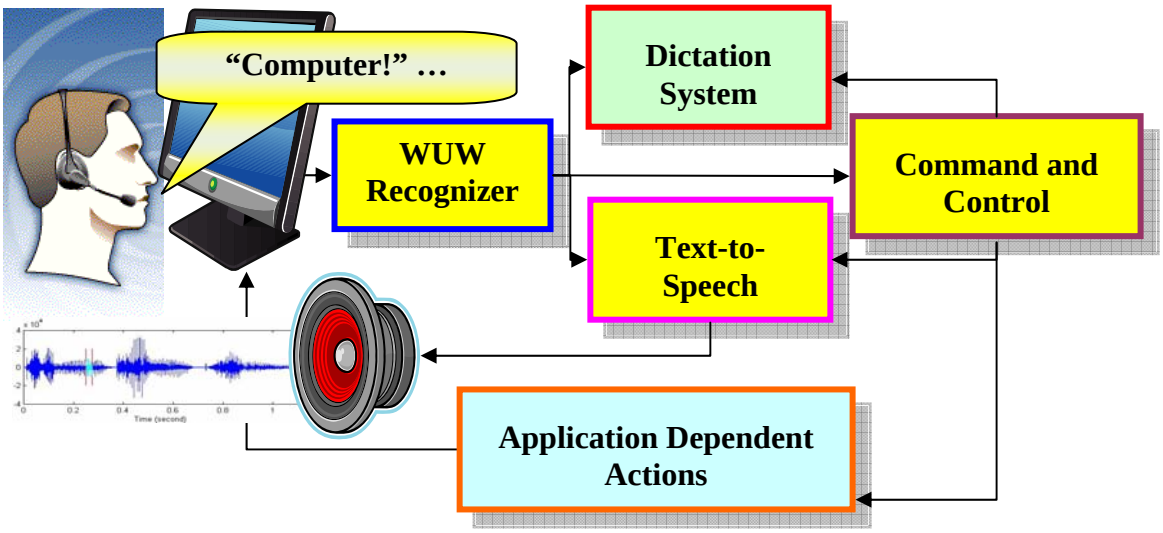
Novel SR technology named Wake-Up-Word (WUW) invented and developed by Këpuska, bridges the gap between natural-language and other voice recognition tasks, Këpuska (2006, 2007, 2008). In order to understand how the system functions it is necessary first to describe this novel paradigm afforded by WUW. *WUW SR is a highly efficient and accurate recognizer specializing in the detection of a single word or phrase when spoken in the **alerting** context of requesting attention, while rejecting all other words, phrases, sounds, noises and other acoustic events with virtually 100% accuracy*. From the presented definition of the WUW paradigm clearly two problems emerge that must be simultaneously solved: (1) Correct WUW detection and recognition – which we call the in-vocabulary (INV) task, and (2) Correct rejection of all other non-WUW's acoustic events – which we call the out-of-vocabulary (OOV) task. In practice

the WUW-SR system must achieve a correct rejection rate of virtually 100% while maintaining high correct recognition rates of over 99% in order to be useful. Since the WUW SR is not yet a fully developed general speech recognition development system, the approach taken for this project is to use WUW SR as a front-end to existing commercial speech recognition development technologies as depicted in Figure 2.

The uC's voice driven interface will support the following tasks: (a) Command-and-Control; (b) Text-to-Speech; and (c) Dictation. The high level software architecture of the system also is depicted in Figure 2. The additional speech technologies that will be used in prototype development are Microsoft's Speech Recognition Development Toolkit (SDK) free with Windows operating systems hardware platforms. In addition other commercial technologies can be licensed if uC must run in platforms other than Windows (e.g., Nuance Inc.).

The reliability of WUW opens up the world of SR to missions that are not achievable with current state of the art SR systems. WUW is markedly different than standard SR algorithms that must use a ***push-to-talk*** paradigm. Current SR applications must resort to the "push-to-talk" paradigm in order to restrict recognition to true-positives (e.g., the users will utter In-Vocabulary – INV utterances only) and removing the possibility that the recognizer will produce false-positive (e.g., Out-Of-Vocabulary – OOV utterances). Thus the operating point of the recognizer is determined by maximizing the discrimination among a finite, and typically small, number of possible choices. WUW recognizer on the other hand performs recognition by considering all possible false-positives while matching and exceeding the correct recognition of INVs compared to state-of-the-art SR technologies.

Figure 2: Software Architecture of WUW-driven general Human-Machine Interactive System.



The centerpiece of this MMUI architecture is robust WUW Speech Recognition technology that enables voice driven context switching - requesting explicitly and implicitly "attention" from the computer (e.g., alerting context) to perform a task (e.g., "**Computer**, Dictation Mode"). The

computer will then be able to switch its actions based on the request. The user will be prompted when necessary again using appropriate MMUI; sound, speech, visual display, etc. This architecture is further clarified in the proceeding section.

Intellectual Merit

One of the goals of the uC system is to allow ***natural*** communication between users and computers via speech, where ***natural*** implies similarity to the ways humans interact with each other on a daily basis. A major obstacle to this is the fact that most systems today still rely to some extent on non-speech interfaces, such as pushing buttons or clicking with a mouse when speech only input is desirable. However, much like a human assistant, a natural speech interface must be ***continuously listening*** and must be robust enough to recover from any communication errors without non-speech intervention.

Speech recognizers deployed in ***continuously listening*** mode monitor acoustic input all the time and do not require non-speech activation. This is in contrast to the ***push-to-talk*** model, in which speech recognition is only activated when the user pushes a button. Unfortunately, today's ***continuously listening*** speech recognizers are not reliable enough due to their insufficient accuracy, especially for ***correct rejection***. For example, such systems often respond erratically, even when no speech is present since they sometimes interpret background noise as speech, and they sometimes incorrectly assume that certain speech is addressed at the speech recognizer when in fact it is targeted elsewhere (context misunderstanding). These problems have traditionally been solved by the ***push-to-talk*** model: requesting the user to push a button immediately before or during talking. Alternatively, the users of such systems are often tempted to control the recognizer by altering the input source, such as muting or removing the microphone between speech recognition sessions, in effect creating another version of ***push-to-talk***. Neither of these situations are representative of a natural speech recognition system.

Another problem with traditional speech recognizers is that they cannot recover from errors gracefully, and often require non-speech intervention. Any speech-enabled human-machine interface based on natural language relies on carefully crafted dialogues. When the dialogue fails, currently there is not a good mechanism to resynchronize the communication, and typically the transaction between human and machine fails by termination. A typical example is a SR system which is in a dictation state, when in fact the human is attempting to use command-and-control to correct a previous dictation error. Often the user is forced to intervene by pushing a button or keyboard key to resynchronize/manually correct the system. Current SR systems that do not deploy the ***push-to-talk*** paradigm use ***implicit context switching***. For example, a system that has the ability to switch from "***dictation-mode***" to "***command-mode***" does so by trying to infer whether the user is uttering a command rather than dictating text. This task is rather difficult to perform with high accuracy, even for humans. The ***push-to-talk*** model uses ***explicit context switching***, meaning that the action of pushing the button explicitly sets the context of the speech recognizer to a specific state. However, to achieve that it uses a different interface modality (i.e., pushing the button).

To achieve the goal of developing a natural speech interface, it is first useful to consider human to human communication. Upon hearing an utterance of speech a human listener must quickly make a decision whether or not the speech is directed towards him or her. This decision

determines whether the listener will make an effort to "process" and understand the utterance. Humans can make this decision quickly and robustly by utilizing visual, auditory, semantic, and/or contextual clues. Thus, understanding how humans express the need for attention is the most important factor that will enable development of intelligent systems that will allow users to communicate more naturally with them. As a result of the authors' research over the past eight years, a robust statistical method capable of recognizing with high accuracy a specific word/phrase, called Wake-Up-Word (WUW), with practically no false alarms was developed. Presently, the only additional acoustic parameter being exploited is a brief leading and trailing silence typically expected in the user's dialog when he/she utters a WUW. While this acoustic clue is backed up with our intrinsic understanding of human behavior it certainly it is not the only way that a human would utter an attention grabbing word/phrase. In contrast, current state of the art commercial SR systems employ additional modalities to direct otherwise voice centered dialog. For example, recently Ford has launched a series of TV commercials depicting their vehicles equipped with intelligent "voice-activated" command system (i.e., SYNC technology developed by Microsoft [Microsoft SYNC]). However, any interaction of the system with the driver is achieved only when the driver presses a special button indicating to the computer his intention to issue a command – a "push-to-talk" paradigm. Clearly, current state of the art SR systems are not suitable for applications requiring continuous monitoring of user input to detect speech directed to the system. In addition to being able to continuously monitor acoustic events those systems must also be trustworthy and effective.

A computational modeling approach to explore the links between prosodic features and their visual counterparts to reveal the semantic information in images and video sequences for realistic language processing in ways useful for human computer/robot interactions is being developed. Visual clues might be gestures such as hands movements or facial expressions.

Wake-Up-Word is proposed as a method to explicitly request the attention of a computer using a spoken word or phrase. The WUW must be spoken in the context of requesting attention and should not be recognized in any other context. After successful detection of WUW, the speech recognizer may safely interpret the following utterance as a command. The WUW is thus analogous to the button in *push-to-talk*, but the interaction is completely based on speech. Therefore the ***explicit context switching*** will be deployed in uC via WUW. During the development and evaluation of this approach we will examine the extent to which users find interaction with the system to be natural and intuitive.

Impact

Successful integration of WUW SR technology will impact not only the uC system but also enable development and deployment of this SR technology in numerous other applications not afforded today such as smart homes, smart court rooms, enabling people with disabilities, smart airplanes, etc., Këpuska and Klein (2008), Këpuska (2006) and Klein (2007), changing the way humans interact with their computerized systems. While commercial speech recognition technology has been around for over 20 years, with WUW technology we can revolutionize human-machine interaction.

AIDING DETERMINATION OF ALERTING CONTEXT WITH INTEGRATION OF VISUAL CLUES

Goal

Humans can easily integrate audio and visual clues from other individuals for various cognitive tasks, for example, for detecting intention or understanding spoken speech. The interaction between hearing and vision in speech perception is known as the McGurk effect (McGurk & MacDonald, 1976). Human speech recognition performance is affected adversely by the presence of acoustic noise; however, this adverse effect is smaller than that observed with audio-only computer speech recognition systems. Earlier works show that utilizing visual information along with audio information improves ASR accuracy, significantly especially under adverse acoustic conditions (Gurbuz, Patterson, Tufekci, & Gowdy, 2001). For humans, it has been shown that the presence of the visual modality is roughly equivalent to a 12 dB gain in acoustic SNR, Sumby and Pollak (1954). Computer speech recognition systems show similar benefits. For example, integration of audio and visual features using a multi-stream state synchronous HMM (MS-SS-HMM) improved word recognition accuracy by more than 45% in our earlier work with test cases of 12 dB acoustic SNR conditions (Gurbuz, Tufekci, Patterson & Gowdy 2002).

This research task's goal is to integrate audio and visual information to distinguish the ***alerting*** context from the ***referential*** context in addition to improving the performance of the multimodal user interface. To the best of our knowledge this has not been studied previously. For this study, we propose a hidden Markov modeling (HMM) based computational approach that infers information from acoustic and visual speech features for a higher level decision to discriminate between an ***alerting*** context and a ***referential*** context. The acoustic features are based on Mel-Frequency Cepstral Coefficients (MFCC) and prosodic features (fundamental frequency (F_0), energy, and duration). The visible features are head pose and gaze information, PCA (principal component analysis) coefficients and affine invariant Fourier descriptors from 3D face data reconstructed by utilizing a pair of off-the-shelf stereo cameras.

Approach

This study will build on our earlier work. We have already developed a novel three-dimensional (3D) face data reconstruction and head pose estimation system that works on a standard personal computer utilizing a pair of off-the-shelf stereo cameras (Gurbuz & Inoue, 2008). The key features are (1) the way the 3D face data is reconstructed by constraining stereo matching using the 3D locations of the eyes, and (2) the determination of a set of face voxels that are relevant for face plane estimation. The former yields an accurate 3D face reconstruction, whereas the latter provides robust head pose estimation. The system is model- and initialization-free, and does not require head attachments or special imaging equipment. Moreover, by not relying on person-dependent features the system minimizes the effect of facial differences among different individuals. The evaluation experiments conducted with a commercial motion capture system

demonstrated that the system is robust and suitable for human-computer interface applications with an average RMS error less than 3.5 degrees. Further, we propose gaze direction be estimated to derive cognitive cues about the user, ranging from gesture information to intention. Visual speech clues will be extracted from the mouth area of the reconstructed face data. Namely, PCA features will be extracted to include the protrusion, lateral and vertical spread-ness information. Further, three-dimensional affine invariant Fourier descriptors extraction algorithm will be derived utilizing the 3D lip contour data to include structural shape information from the tracked lip contours, Gurbuz (2005). These two sets of features along with their derivates will represent visual speech information and be integrated with the audio information.

Speech intelligibility differs significantly depending on the noise spectral energy distribution even at the same SNR levels (e.g., car noise is less destructive than factory noise at 6dB SNR level). Our earlier experimental results highlight that audio reliability is closely correlated with SNR, but SNR itself is not a sufficient measure for audio reliability for multimodal speech recognition systems. Further, our earlier results suggested that audio stream should be weighted dynamically (determined from the signal) based on the effect of the corrupting noise, and a multi-stream state synchronous integration method appears to be a suitable framework for such an on-the-fly weighting scheme. For multi-stream integration, it is best to utilize multi-audio Mel-frequency filter bank channels for their individual stream weighting based on the spectral energy information of the interfering noise and SNR level. Therefore, we will develop an algorithm to investigate how ambient acoustic noise degrades the spectrum of human speech for developing highly effective adaptation mechanisms for the integration of audio and visual speech information for improving the speech recognition systems.

An audio-visual multi-stream state synchronous HMM architecture can be formed by modification of an existing audio only HMM architecture without needing a re-training of the audio HMM models and without a large audio-visual database. That is, the video HMM models can be trained from a smaller audio-visual data set and then the visual models will be combined with the audio HMMs trained on a large data set to form state synchronous audio-visual HMMs. This structure will be utilized for audio-visual speech recognition as well as determining an ***alerting*** context from a ***referential*** context to provide means to interact with the system in a natural manner.

For this task, we will be preparing a database for various test cases including emotion detection and gaze tracking. This would evolve into a comprehensive database for human-computer interface (HCI) research. This as well as the code for any tools developed will be made available to the research community.

Intellectual Merit

One of the goals of the uC system is to allow ***natural*** interaction between users and the system via a noise robust multimodal user interface. The study will reveal how ambient noise corrupts human speech and its adaptation method within the audio-visual speech recognition framework. This study will also reveal an efficient method for multimodal data fusion. Further, this study will explore the links between acoustic features and their visual counterparts to reveal the

semantic information in images and video sequences for intention detection and language processing in ways useful for human computer interactions.

Impact

The outcomes of this research will move HCI research one step forward and provide the means for applications (humanoid robots, computers and smart homes) to interact with users in a natural manner. Moreover, the study of ambient noise interference on human speech may lead to a better noise cancellation methodology for improved cochlear implant devices and hearing devices.

INTEGRATE VISUAL SIMULATION FUNCTIONALITY INTO UC PROTOTYPE

Goal

In this task the uC researchers will develop additional visual simulation techniques to augment the capabilities already developed. The new capabilities apply to a broad range of real-world environments. These include not only normal complex collaborative tasks (e.g., construction), but also situations with limited visibility, such as mining, nuclear power plants and underwater recovery operations, where onsite visualization tools can substantially aid collaborative problem solving and decision making. Thus, they provide fertile ground on which to support collaborative problem solving.

Approach

We have already developed a uC visual simulation capability (3-D viewer) to improve the perception and understanding of scenes where near real-time data is available. Algorithms, heuristics and software development lessons learned from implementing these complex systems will be applied to the new uC development efforts. The uC's 3-D viewer architecture consists of three families of network-enabled applications and services: data distribution, data acquisition, and data visualization. The core of the data distribution suite consists of a real-time database server and a publish-and-subscribe service library. The real-time database server is responsible for maintaining an accurate representation or world model of all the elements that compose the scene. The publish-and-subscribe library allows all other applications to synchronously and concurrently receive update notifications and query information about the world model. The data acquisition suite consists of applications customized to gather data from specific sources and publish the information to the real-time database server. This suite of applications also includes database access stubs and general-purpose simulators. Together, the data acquisition applications are responsible for updating the world model so that it accurately represents the scene.

Impact

Although the description above is oriented to harsh construction environments, it is also applicable to other situations where incomplete or dynamic topological information may be available with respect to the environment. For example, the proposed system can be used in large warehousing, healthcare and construction operations to make routing decisions in 3-D. In addition, this visually-based decision-making system will be applied to other fields such as

emergency management, aviation, military, ship-building and tracking, service, manufacturing, construction and underwater searches (Rodriguez, Opdenbosch & Santamaria, 2006).

INTEGRATE NETWORKED SENSORS FEEDBACK INTO UC PROTOTYPE

Goal

Central to uC team performance effectiveness is its ability to push and pull information from various networked sensors. We will develop and use sensor technologies as identified by their function and connectivity: (a) Sensors that are directly connected to (most likely handheld) uC devices; such as, cameras or an array of microphones as well as sensors specialized for specific use of the device (e.g., cardiac pulse monitoring sensor); and (b) Sensors connected to the network accessible by uC; such as, Accelerometers, Pressure Sensors, Gyroscopes, Piezoelectric Sensors, Geophones, Microphones, Cameras. The function of the sensors connected to the uC groups them as a sensor that serves the purpose of: (a) Controlling the device itself and (b) Sharing its data with other team-learners connected to the network. Note that a sensor may have dual use; for example, an array of microphones can be used to control the device (e.g., "uC, call ***Cynthia!"***) as well as share the data with users (i.e., voice is being transmitted over the network).

Approach

Integration of a large variety of sensors producing distinctive data measurements can be achieved within existing WWW technologies. Namely XML in conjunction with XSL is specifically designed to bridge the gap of heterogonous data representation. XML is a general-purpose markup language. It is used to facilitate the sharing of structured data across different information systems (i.e., Internet). It allows definition of custom tags. XSL is a language for expressing style sheets. An XSL style sheet is a file that describes how to display an XML document of a given type. To achieve this XSL contains: **XSLT:** A transformation language for XML documents. It is used as a general purpose XML processing language. XSLT is thus widely used for purposes other than XSL, like generating HTML web pages from XML data. In uC this will allow us standardization of the displaying software namely use of browsers; **XPath:** A language used for navigating in XML documents; **XSL-FO:** Advanced styling/formatting features, expressed by an XML document type which defines a set of elements called **Formatting Objects**, and attributes. Special consideration will be given to capabilities of current technologies (Cullom & Sedbrook, 2007; Lee et al., 2003) as well as hardware-software architectures of the uC supportive systems (Sandman & Riley, 2003).

Intellectual Merit

(1) uC will explore the acquisition of real-time data from embedded sensors and feed that data into the appropriate problem solving challenge; (2) it will do this in the context of team-task roles and their interdependencies (i.e., sensors help populate and make more meaningful the context in which the team is interacting---whether it be 3D, 2D, voice, its transcription, or action triggered by it); (3) the sensors augment reality such that real-time information (e.g., storm surge height/locations) is inextricably linked with the visual simulation - the simulation is not based upon pre-canned models; (4) while some systems help people collaborate online, none of them

have functionality that infuse real-time environmental "knowledge" into that interaction in a way consistent with the principles of team/task work; and (5) none use context-based speech or multi-modal recognition and visualizations. Building on artificial intelligence, Bell, Feiner and Höllerer (2002), and team cognition research, Salas anf Fiore (2004), the uC system will tag scene items and convey related content to online team-learners and collaborators. Further, this work will be extended by the inclusion of practitioners' wisdom as well as timely, contextual data, information and knowledge.

Impact

Gathering real-time data in this fashion will facilitate sense-and-response decision-making, problem-solving and design issues relevant to numerous collaboration needs in industry and education. These uC capabilities are unique and bring about team performance niche market and commercialization opportunities. Thus, the broader impact resulting from the proposed activity is a robust and open technology based system and infrastructure that will support high performing teams in complex problem solving and decision making environments.

INTEGRATE TEAM PERFORMANCE SCAFFOLDING INTO UC PROTOTYPE

Goal

uC's process scaffolding capabilities will provide tele-presence and mobile technology tools to augment dynamic problem solving and decision making. By incorporating real-time data from both distributed sensor systems and team members dispersed throughout the globe, with group support software, this element of uC will result in significant gains in effectiveness across industries as diverse as biotechnology, nanotechnology, software design, construction and many engineering fields. Given that collaborative problem solving often suffers from process break-downs, with this component of uC we overlay a theoretically grounded understanding of team-process. Our goal is to scaffold the interaction processes in support of effective problem solving when addressing complex, real-world tasks.

Approach

The backdrop against which the uC engine is developed is the notion of *team competencies* - factors that foster effective interaction behaviors and performance (Cannon-Bowers, Tannenbaum, Salas, & Volpe, 1995). Our research will pursue innovative mobile and fixed location uCollaborator technologies based upon the team competencies framework. Table 2, adapted from Cannon-Bowers et al. (1995), illustrates this important point and includes examples of how uC can support foundational team processes. Specifically, we will focus on creating systems that are able to utilize real-time data from team members who are not co-located and from sensors in the field to support distributed interaction as collaboration unfolds dynamically. First, through the development of simulation-based visualization and group support software, uC will provide a powerful range of collaboration tools usable in team problem solving situations. Second, through the use of hand-held and mobile devices, our technologies will support collaboration with distributed members who may not have access to high-end simulations.

Table 2: Team and Task Factors Supported Through uC.

Relation to Team	Relation to Task	
	Specific	Generic
Specific	*Context driven* – describes for example, data provided through task sensors and team members in the field helping to understand nature of problem/decision requirement	*Team-contingent* – describes, for example, support of processes related to characteristics of the team such as structure, roles
Generic	*Task-contingent* – describes, i.e, data pertaining to factors such as task sequencing and development of mental models relevant to problem/decision requirement	*Transportable* – describes, for example, support of processes such as planning or effective communication

Representative generic team and task factors that would be supported include conflict resolution (often a problem with distributed team projects), collaborative problem solving, communication, performance management and planning and task coordination. For example, a mobile component of the uC system would scaffold planning processes via support of information management to align team interdependencies (e.g., real-time data targeting team leaders). A fixed component of our system would use simulations to scaffold collaborative problem solving – that is, simulations to help team members identify critical problem cues and effectively represent such data in service of eliciting appropriate team member participation. Creating and implementing such a system provides a unique opportunity through which to support a tremendous variety of complex collaborative tasks – impacting a number of industries and leading to significant economic development and increased productivity.

Impact

The societal impact here rests on the fact that problems in coordination and communication continually create financial losses. For example, uC can be used not only to help redress design problems before they occur in areas such as construction, supply-chain, or software development, but our concept also holds great promise in enabling synchronization of complex efforts involving multiple teams. With regard to financial losses, as noted in a recent industry white paper, "the cost of poor quality – not doing it right the first time – in the software industry today is an estimated $100 billion. Seventy five percent (75%) of software projects undertaken are never completed. (Source: http://www.wrsystems.com/whitepapers/ qatesting.pdf.) The context of interacting in a multiple team setting is made even more challenging by virtue of the fact that success requires rapid transfer of information both within the team, and also across the boundaries of other teams with whom teams may or may not have any prior experience working. Systematic R&D in uC based technology would significantly impact coordination of teams by digitizing and rapidly transmitting information regarding the

status of each team, transferring newly acquired information to other units, and enabling teams to perform distinct aspects of their tasks while properly supporting the efforts of other team members or other teams in the system.

EVALUATION OF UC-USER INTERACTION

Goal

As noted previously, in order for the uC system to be of substantial practical value in facilitating team performance, its components must be designed and integrated such that users find the system to be simple and intuitive, even when interacting with its most complex features. Without careful attention to user behavioral tendencies and preferences, the full potential of the system may not be realized. Thus, assessment of uC-user interactions will be an integral part of each step in the development of the system, informing the design of each element for improved usability.

Approach

Given evidence indicating that the manner in which humans communicate with computers tends to be different from how they communicate with other humans (e.g., see Dahlbäck, Jönsson, & Ahrenberg, 1998), addressing the first issue will involve assessing and recording how individuals would naturally interact with the various system components (e.g., their speech patterns and head position when attempting to gain the attention of the system). This information will then be used in designing and refining the system to allow for intuitive user-system interaction. To this end, a Wizard-of-Oz approach will be employed during the initial stages of system development within a controlled laboratory setting. In this approach, team members will be recruited to participate in experiments in which they interact with the system at different stages of development. During the initial stages, participants will be informed of the relevant system components with which they will work (e.g., speech recognition for certain tasks); however, system behavior will be controlled by an experimenter rather than the system itself (with participants unaware of this). This will allow us to capture natural behavioral tendencies when interacting with that aspect of the uC system (e.g., speech patterns when attempting to gain the attention of the system). During the later stages, participants will be informed of the relevant system components with which they will work; however, at these stages system behavior will in fact be controlled by the system itself. This will allow us to evaluate and improve system recognition accuracy.

To address the second issue, usability testing will be conducted in a real-world setting. Consistent with recommendations for project evaluation (e.g., Frechtling, 1997, 2002), this assessment will involve collecting information using both qualitative and quantitative approaches to identify how the uC is adopted and used by people in their working and everyday lives. Such settings are very different from the laboratory environment used during the earlier Wizard-of-Oz approach, where tasks are set and completed in an orderly way. The real world usability study will enable researchers to observe the uC in a more realistic or "messy" context in the sense that activities often overlap and are constantly interrupted. Through the evaluation of how people think about, interact, and integrate the use of the uC within the setting it will ultimately be used, a better sense of how successful the products will be in the real world will be identified. In this case, qualitative data focusing on accounts and descriptions of people's behavior and activities

with using the uC will be obtained. Data will be collected primarily by observing and interviewing people; collecting video, audio, and field notes to record what occurs in the chosen setting.

Intellectual Merit

Distributed teamwork continues to impact competitiveness as collaboration across time and space becomes increasingly frequent in the workplace and the edu-space. The evaluation task through usability evaluations resulting in continuous design recommendations will optimize the potential for team performance and ubiquitous team-collaboration; that is, collaboration linking both co-located and geographically distributed people and data, information and knowledge. The evaluation task will specifically uncover the limitations and potentials of this system as it applies to team collaboration.

Impact

The evaluation task will not only consist of conducting usability evaluation in a laboratory setting, but also in a real-world setting to ensure the usefulness of the uC as well as a user-centered focus in terms of design of the speech-recognition component within the uC resulting in revolutionary impacts in the field of distributed team performance. Furthermore, the results from the usability evaluation studies will have practical implications and theoretical results of broad importance to the development of the uC enabling its use across multiple settings and within several industries

REFERENCES

Cisco TelePresence: http://www.cisco.com/en/US/solutions/ns669/expert_on_demand.html

Cullom, C. & Sedbrook, T. (2007). Capturing and Shaping Shifting Requirements using XML and XSLT: A field Study, *Journal of International Technology and Information Management,* 16(1), 73-86.

Dahlbäck, N., Jönsson, A., & Ahrenberg, L. (1998). Wizard of oz studies – Why and how. In M. T. Maybury & W. Wahlster (Eds.), *Readings in Intelligent User Interfaces* (610-620). San Francisco, CA: Morgan Kaufmann

Frechtling, J. (1997). *User-friendly handbook for mixed method evaluations*. Darby, PA: DIANE Publishing Company**.**

Frechtling, J. (2002). *The 2002 user friendly handbook for project evaluation* (REC 99-12175 evaluation of NSF educational programs). Washington DC: National Science Foundation.

Freeware Palm: *Speech* and Debate Timekeeper v2.2*: http://www.frewarepalm.com/clock/*

Gurbuz, S., Patterson, E. K., Tufekci, Z., & Gowdy, J. N. (2001). Affine-Invariant Visual Features Contain Supplementary Information to Enhance Speech Recognition, *International Conference on Audio-and-Video-Based Biometric Person Authentication,* Sweden, 2001.

Gurbuz, S., Tufekci, Z., Patterson, E. & Gowdy, J. N. (2002). Multi-stream product modal audio-visual integration strategy for robust adaptive speech recognition,. in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2002.

Gurbuz, S. (2005). Real-time Outer Lip Contour Tracking for HCI Applications, *Interspeech 2005 - European Conference on Speech Communication and Technology, Lisbon, Portugal*, September, 2005.

Gurbuz, S., & Inoue, N. (2007). Real-Time Head Pose Estimation Using Reconstructed 3D Face Data from Stereo Image Pair, *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Hawaii, USA.

Huang, X., Acero, A. & Hon, H. (2001). Spoken Language Processing: A Guide to Theory, Algorithm and System Development. Prentice Hall PTR.

Këpuska, V. (2006a). Dynamic Time Warping (DTW) Using Frequency Distributed Distance Measures: US Patent: 6983246, January 3, 2006.

Këpuska, V. (2006b). Scoring and Rescoring Dynamic Time Warping of Speech**:** US Patent: 7085717, April 1, 2006.

Këpuska V. (2006c). Wake-Up-Word Application for First Responder Communication Enhancement, SPIE, Orlando, 2006.

Këpuska, V., Carstens, D. S., & Wallace, R. (2006, March) Leading and Trailing Silence in Wake-Up-Word Speech Recognition, Proceedings of the International Conference: Industry, Engineering & Management Systems 2006, Cocoa Beach, FL., 259-266.

Klein, T. (2007). Tripple Scoring of Hidden Markov Models in Wake-Up-Word Speech Recogntion, Thesis, Florida Institue of Technology.

Këpuska, V., & Klein, T. (2008). On Wake-Up-Word Speech Recogntion Task, Technology, and Evaluation Results against HTK and Microsoft SDK 5.1. Invited Paper: World Congress on Nonlinear Analysts, Orlando 2008, To appear in Journal of Nonlinear Analysis, Theory, Methods & Applications.

Lea, B., Yu, W. & Kannan, P., (2007). Social Network Enhanced Digital City Management and Innovation Success: A Prototype Design, *Journal of International Technology and Information Management,* 16(3), 1-22.

Lee, H., Etnyre, V., & Chen, K. L. (2003). A Study of .Net Framework, XML Web Services and Supply Chain Management, *Journal of International Technology and Information Management,* 12(1), 137-153.

McGurk, H. & MacDonald, J. (1976). Hearing lips and seeing voices, *Nature*, 264(5588), 746–748.

Metcalf, D. S. (2007a). Innovations in eLearning. George Mason University. Fairfax, VA. June 6-7, 2007.

Metcalf, D. S. (2007b). eLearning Guild Annual Gathering. Boston, MA. April 10-13, 2007. Topic: ISD for mLearning, Learning Technology Innovations: Mobile, Wikis, Podcasts and More....

Metcalf, D. S. (2007c). mLearning Innovations. Webinar. March 29, 2007. Academic Impressions.

Milone, D. & Rubio, A. (2003). Prosodic and accentual information for automatic speech recognition. *IEEE Transactions on Speech and Audio Processing*, 11(4), 321-333.

Pinker, S. (1994) *The Language Instinct: How the Mind Creates Language.* New York: HarperCollins.

Rodriguez, W., Opdenbosch, A. & Santamaria, C. (2006). Managing Construction Operations Visually: 3-D Techniques for Complex Topography and Restricted Visibility, *ASEE Engineering Design Graphics Journal*, 70(2), 6-15.

Salas, E., & Fiore, S. M. (Editors). (2004). Team Cognition: Understanding the factors that drive process and performance. Washington, DC: American Psychological Association.

Sandman, T., & Riley, T. (2003). Selecting Middleware For N-Tier Applications, *Journal of International Technology and Information Management,* 12(1), 107-120.

Sumby, W. H. & Pollak, I. (1954). "Visual contributions to speech intelligibility in noise, *Journal of Acousics of the. Society of America*, 26(2), 212–215.

Zue, V. & Cole, R**.** (1997). Survey of the State of the Art in Human Language Technology. Cambridge University Press and Giardini.