Efficiency of Speech Recognition for Using Interface Design Environments by Novel Designers

MOHAMMAD M. ALSURAIHI & DIMITRIS I. RIGAS
Department of Computing
University of Bradford
Bradford, West Yorkshire, BD7 1DP
ENGLAND

Abstract: - Previous studies on usability of graphical design-widgets, like menus and buttons, proposed the use of speech and non-speech (earcons and auditory icons) for solving their usability problems. In this paper we investigate speech as an input metaphor to enhance learnability, or the ability to use a system with no prior knowledge, in order to design interfaces using a multimodal interface design toolkit called MMID. Using this toolkit as an experimental platform, the paper presents an empirical multi-group study that compares efficiency of visual-only and multimodal interaction metaphors when used by novel users.

Key-Words: - interface-design, usability, learnability, experienced performance, efficiency, multimodal interaction, voice-instruction, speech.

1 Introduction

The heavily emphasis on visual interaction in the currently existing toolkits to design interfaces is making interaction systems more and more visually crowded. This makes interfaces become 'nervous' and 'oppressive' to the user [30], and causes the user to experience information overload [6, 31] by which important information may be missed [39]. Another problem with visual-only interaction is the high potential for usability problems with graphical metaphors to occur. There are two root-problems from which all usability problems branch: interface intrusion into task [9, 35, 36] and closure [7, 17, 18].

The key solution to enhancing usability of graphical interfaces is to lessen the visual workload on the visual channel, which negatively affects efficiency of task-performance [10, 55]. There is a growing body of research that recommends the addition of non-speech sounds (earcons [3, 5, 8, 9, 29, 38, 43, 52-54] and auditory icons [11, 12, 20-24, 57]) to interfaces in order to improve their performance and increase their usability. However, in order for perception and right interpretation of non-speech sound to be successfully achieved, a high level of concentration and the development of a perceptual context are required by the users [27, 48-53]. This causes the users to incorrectly interpret the musical messages sometimes, because of lack of concentration and distraction with other events or messages in the interface [53]. Processing natural language (speech) feedback in the interface has been recognised as offering many benefits in human-computer interaction [32-34]. Rigas et al [44-47, 53] investigated the use of speech along with non-speech sounds, and found that the combination of earcons along with synthesised speech was a successful and effective approach for communicating information to the users. In addition, a study by Vargas and Anderson showed that the users' performance was better in terms of time, number of keystrokes, errors, and workload when used speech along with earcons [56].

The work by Bolt [4], which introduced the approach of processing speech and gesture for moving graphical objects on an interface, was pioneer and promising for researchers to investigate the use of speech as an input and output utility for enhancing efficiency of the user-interface. The studies by Cohen, Oviatt and others [11, 13-15, 37, 40, 41] recommended the employment of speech recognition for utilizing the user interface. There is a worry that interacting with the interface vocally would not be as efficient as interacting with it visually, because of recognition errors [13, 25]. However, these errors are tameable and can be tolerated [42, 58], especially if interaction can be enhanced with limited amount of vocabulary [28]. Previous work on speech as an input means has shown its potential for enhancing performance of

interaction between the user and the interface [1, 2, 16, 19, 26]. In this paper we investigate efficiency of vocal instruction against visual-only instruction for interface design with consideration to limit the use of the common graphical metaphors used in IDEs like menus, toolbar, toolbox, properties-table and status-bar.

2 Experimental Toolkits

Two experimental toolkits were developed using Microsoft Visual C#: Typical Visual-Only Interface Design (TVOID) and Multi-Modal Interface Design (MMID) toolkits. TVOID imitates the style of interaction implemented in most of the existing interface-design environments like Microsoft Visual C# and Java NetBeans IDE. It interacts with the user visually-only with no involvement of other senses like the auditory system. This interaction takes place in six areas in its main interface: menus, toolbar, toolbox, workplace, properties-table, and status-bar. Fig.1 shows a screenshot of TVOID.

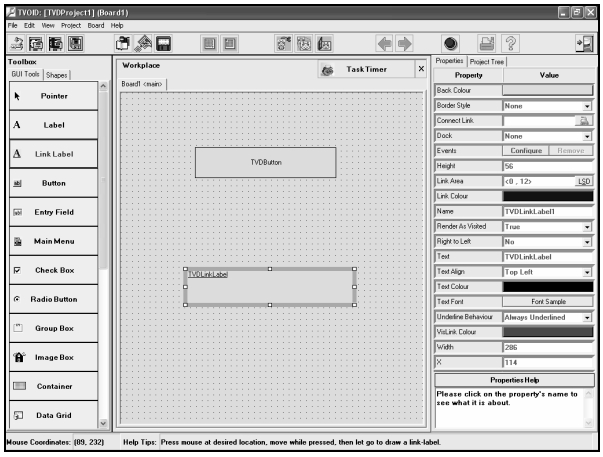

Fig.1: A screenshot of TVOID's main interface

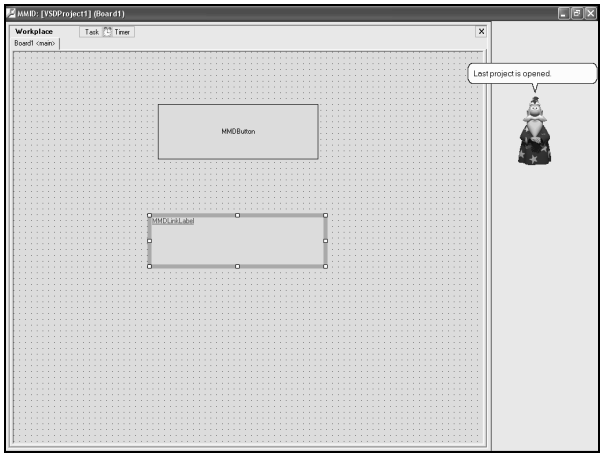

Fig.2: A screenshot of MMID's main interface

MMID provides a combination of visual, vocal and aural interaction metaphors. It is a speech-recognition and text-to-speech based environment that allows limited use of the mouse and the keyboard. It allows the user to interact with it from the position of the mouse-cursor. In this environment, there is no need for the user to use any of the graphical metaphors implemented in TVOID. The system command receptor in this environment is represented by a friendly character (MS Agent) that listens to commands and interacts with the user via speech and facial expressions. Vocal commands are in the form of simple one to three English-words. Fig.2 shows a screenshot of MMID.

3 Experimental Design

The empirical study aimed at measuring efficiency of learnability (or first time use) of the two environments (TVOID and MMID). Efficiency was measured by timing function-learning and task completion, and calculating the number of errors. The toolkits were tested independently by two groups of users (Group A and Group B). Each group consisted of 15 users. The participants were computer users who had limited experience in using interface design environments. Each group was asked to complete 10 tasks. Each task consisted of one to three functions. The tasks were designed to be increasing in complexity and covering all expected functionality (activating menu-command functionality, selecting tools, drawing objects, and setting properties).

4 Results and Discussion

4.1 Task Accomplishment Time

During the experiments, it was noticed that the users who tested TVOID (Group A) expected how to do most of the functions. This environment looked familiar to them because they had previous experience with similar environments that provided with similar interaction metaphors. This experience made them primarily rely on their memory. Before doing a task, the users of TVOID spent time on recalling how to do functions in the similar systems they were used to, to be able to do them using this environment. Expectations of how to do functions were incorrect sometimes, which caused the users to explore how to do these functions. In this way, the users of TVOID did two things to learn functions: remembering or expecting, and exploring in case of incorrect expectation. This was not the case with the users who tested MMID (Groups B). These users were not familiar to voice-instruction and thus they headed directly to exploring.

Fig.3 demonstrates that the users in Group B (MMID group) learned the vocal commands that

replaced the typical textual menu-items faster than their counterparts in Group A who used the graphical menus. This result was mostly the same for learning the other functions as can be seen from Fig.4 and Fig.5. Gathering all commands in one location (e.g. one list) as in MMID helped the users in Group B to locate the required commands for doing functions in a faster way than their counterparts in Group A who looked for these commands in different locations around the interface. Also, the use of one interaction metaphor (voice-instruction) in MMID saved the time for the users to think where to go to activate the required command because of the many encapsulated widgets in the interface (menus, toolbar, drawing tools, properties, etc).

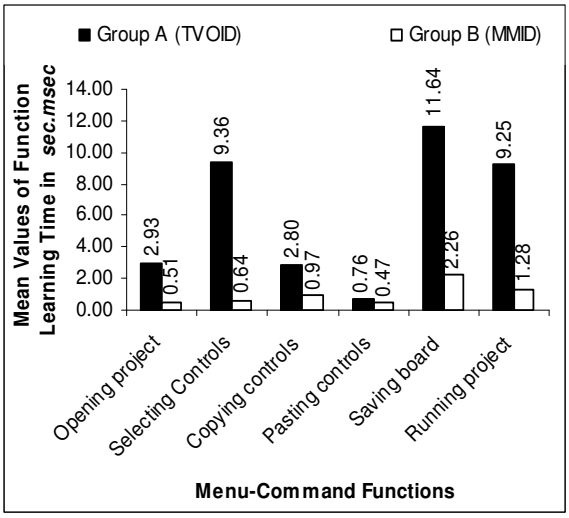

Fig.3: Mean values of time taken by the users to learn 6 menu-command functions using TVOID and MMID

In order to draw an object in TVOID, it must be selected at first, while in MMID the user can draw directly by saying the tool's name on the required location. Specifying mouse coordinates is also easier in MMID as it offers an interactive mouse-cursor that shows the coordinates while moving on the form being designed, while in TVOID the user has to look for them in the status bar. These features facilitated learning these functions as can be seen from Fig.4.

Learning how to set properties was done through interactive training simulations in MMID, while in TVOID was done textually. Enabling the user to specifically learn what he/she needs to learn using interactive training in MMID saved the time for thinking of the appropriate keywords, looking for them, and reading about them as in TVOID. Fig. 5 shows this result. The figure also shows that the alignment property took very long time to learn by most of the users in Group A. The reason of this was mainly because of the name of the property, which was ambiguous. As in Visual Studio .NET, the alignment property in TVOID was called *Dock*. Most of the users could not realise that this word was indicating to the process of aligning objects to sides in the form being designed as most of the users were not speaking English as the first language.

Fig.6 shows the variances between the two environments in regard to task accomplishment (learning and completion). It can be seen that the variances are notable. These variances happened mainly because of the variances in learning functions. Another factor behind this result was that MMID limited the use of the mouse and decreased the reliance on the visual sense. Significance of the difference was tested using the *t*-test. The difference was found significant ($t = 2.64$, $P = 0.02$).

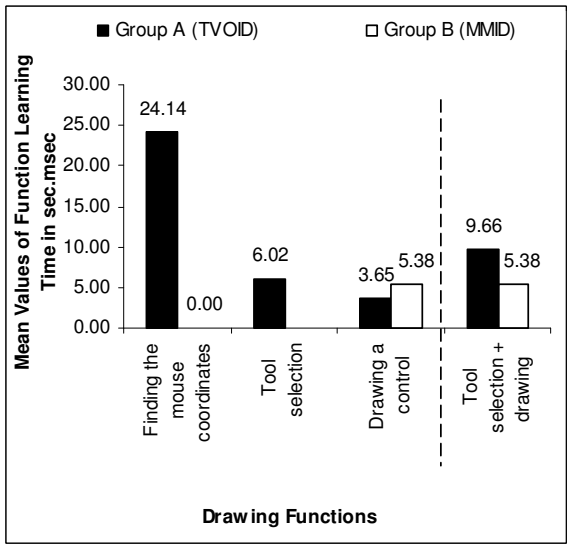

Fig.4: Mean values of time taken by users to learn 3 drawing functions using TVOID and MMID

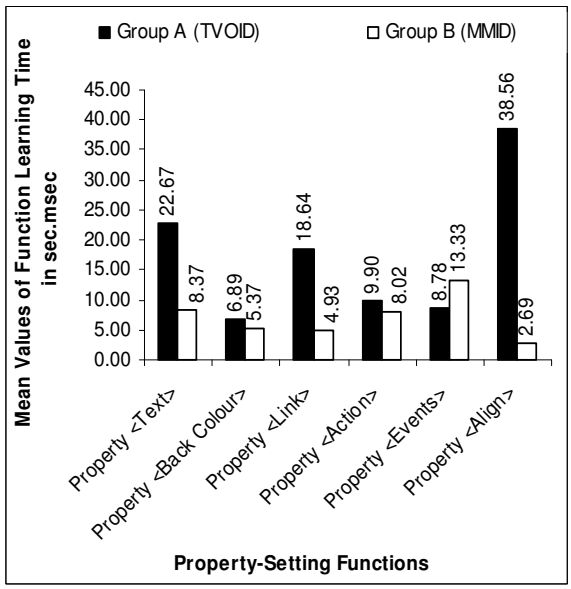

Fig.5: Mean values of time taken by users to learn 6 property-setting functions using TVOID and MMID

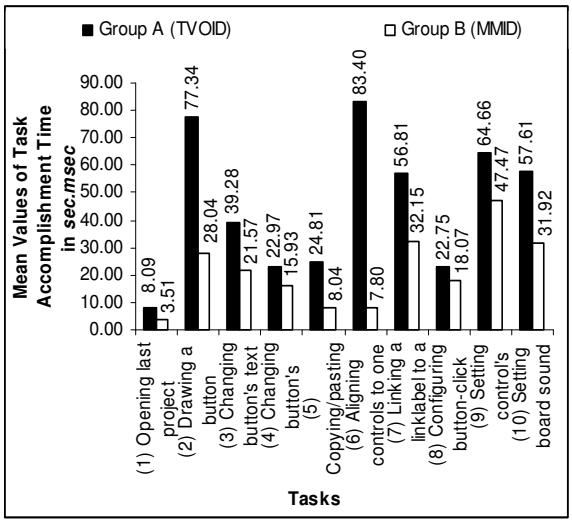

Fig.6: Mean values of time taken for accomplishing 10 tasks for the first time using TVOID (Group A) and MMID (Group B)

4.2 Errors

Calculating the number of errors showed that the users in Group B did more errors than the users in Group A during accomplishment of the same tasks. Table1 shows the frequency of errors made by the users during task accomplishment.

Tasks		Number of Errors	
		TVOID (Grp. A)	MMID (Grp. B)
1	Opening last project	0	3
2	Drawing a button	1	4
3	Changing button's text	1	7
4	Changing button's background colour	0	2
5	Copying/pasting controls	0	2
6	Aligning controls to one side	4	3
7	Linking a link-label to a URL	0	6
8	Configuring button-click action	0	1
9	Setting control's interactive events	9	6
10	Setting board sound events, saving board, and running project	1	7
Total		16	41

Table 1: Frequency of user errors during using TVOID and MMID

The difference between the two environments in regard to number of errors made was found significant (t = 2.17, P = 0.04). Errors in MMID mostly happened because of the recognition problems that occur because of sensitivity toward noise (other voices around) and accurate pronunciation. One of the factors that cause command misrecognition is the lack of confidence in the user's voice when saying the command. The user is afraid that the system will not recognise what he says and thus his/her voice sounds with less confidence. It was noticed that most of the users in Group B had experienced this problem.

In spite of this result, it was noticed that making errors during using MMID made the users more used and familiar to voice-instruction. We anticipate that frequent use of the technology will cause the number of errors during accomplishing tasks to be reduced.

5 Conclusion and Future Work

To shorten task accomplishment time and, hence, enhance the efficiency of task accomplishment, a design environment should aim at enabling the user to do all actions from inside the workplace area with no need to leave it to other areas like menus, toolbar, toolbox, properties-table, and status-bar.

The more visual interaction metaphors an environment provides the more time will be spent in thinking where to find the appropriate ones for accomplishing jobs, and the vice-versa is correct.

The use of voice-instruction as a way of interaction was found to be more efficient than the use of several visual interaction metaphors, in terms of shortening function learning and task-accomplishment time. This study supports the idea of substituting most of the common graphical widgets with other modalities, voice instruction namely, to enhance the usability of interface design environments.

Although MMID was more prone to errors than TVOID because of sensitivity toward noise and accurate pronunciation of words, it must be recalled that it was tried for the first time and that frequent use could lessen the number of errors and make it more usable.

The empirical work covered in this paper investigated the efficiency of two design environments (TVOID and MMID) from one angle, which is learnability or the ability to accomplish tasks from first time use. Further work is needed for testing the Experienced User Performance (EUP) of task-completion in the two conditions (visual-only and multimodal).

References:

[1] Adler, A. and Davis, R., "Speech and Sketching for Multimodal Design", In International Conference on

Computer Graphics and Interactive Techniques, Boston, Massachusetts, pp. 214-216, 2006.

[2] Begel, A. and Graham, S. L., "An Assessment of a Speech-Based Programming Environment", In Visual LAnguages and Human-Centric Computing (VL/HCC'06), pp. 116-120, 2006.

[3] Blattner, M. M., Sumikawa, D. A., and Greenberg, R. M., "Earcons and Icons:Their Structure and Common Design Principle", Human Computer Interaction, vol. 4, pp. 11-44, 1989.

[4] Bolt, R. A., "Put-that-there: Voice and Gesture at the Graphics Interface", In proceedings of the 7th annual conference on Computer Graphics and Interactive Techniques Seattle, Washington, USA, pp. 262-270, 1980.

[5] Brewster, S. A., Wright, P. C., and Edwards, A. D., "A Detailed Investigation into the Effectiveness of Earcons", In proceedings of ICAD '92, pp. 471-498, 1992.

[6] Brewster, S. A., "Using Non-Speech Sound to Overcome Information Overload", Displays, vol. 17, pp. 179-189, 1997.

[7] Brewster, S. A. and Crease, M. G., "Correcting Menu Usability Problems with Sound", Behaviour & Information Technology, vol. 18, pp. 165-177, 1999.

[8] Brewster, S. A., "Sonically-Enhanced Widgets: Comments on Brewster and Clarke, ICAD 1997", ACM Transactions on Applied Perception (TAP), vol. 2, pp. 462-466, 2005.

[9] Brewster, S. A. and Clarke, C. V., "The Design and Evaluation of a Sonically Enhanced Tool Palette", ACM Transactions on Applied Perception (TAP), vol. 2, pp. 455-461, 2005.

[10] Brown, M. L., Newsome, S. L., and Glinert, E. P., "An Experiment into the Use of Auditory Cues to Reduce Visual Workload", In proceedings of ACM CHI '89, Austin, Texas, USA, pp. 339-346, 1989.

[11] Cohen, M. and Ludwig, L. F., "Multidimensional Audio Window Management", International Journal of Man-Machine Studies, vol. 34, pp. 319-336, 1991.

[12] Cohen, M., "Throwing, Pitching and Catching Sound: Audio Windowing Models and Modes", International Journal of Man-Machine Studies, vol. 39, pp. 269-304, 1993.

[13] Cohen, P. R. and Oviatt, S. L., "The Role of Voice Input for Human-Machine Communication", In proceedings of the National Academy of Sciences, pp. 9921-9927, 1995.

[14] Cohen, P. R., Johnston, M., McGee, D., Oviatt, S. L., Pittman, J., Smith, I., Chen, L., and Clow, J., "QuickSet: Multimodal Interaction for Distributed Applications", In proceedings of the Fifth ACM International Multimedia Conference, New York, pp. 31-40, 1997.

[15] Cohen, P. R., Johnston, M., McGee, D., Oviatt, S. L., Clow, J., and Smith, I., "The Efficiency of Multimodal Interaction: A Case Study", In proceedings of the International Conference on Spoken LAnguage, Sydney, Australia, 1998.

[16] Desilets, A., Fox, D. C., and Norton, S., "Voce Code: An Innovative Speech Interface for Programming-by-Voice", In Extended Abstracts of the 2006 Conference on Human Factors in Computing Systems (CHI 2006), Montereal, Quebec, Canada, 2006.

[17] Dix, A., Finlay, J., Abowd, G., and Beal, R., Human-Computer Interaction, Second ed: Prentice Hall, 1998.

[18] Dix, A. J. and Brewster, S. A., "Causing Trouble with Buttons", In Ancillary Proceedings of BCS HCI '94, Glasgow, Scotland, 1994.

[19] ElAarag, H. and Schindler, L., "A Speech Recognition and Synthesis Tool", In proceedings of the 44th Annual Southeast Regional Conference, Melbourne, Florida, pp. 45-49, 2006.

[20] Fernstorm, M. and McNamara, C., "After Direct Mainpulation---Direct Sonification", ACM Transactions on Applied Perception (TAP), vol. 2, pp. 495-499, 2005.

[21] Gaver, W., "The SonicFinder: An Interface that Uses Auditory Icons", Human Computer Interaction, vol. 4, pp. 67-94, 1989.

[22] Gaver, W., "Auditory Interfaces", in Handbook of Human Computer Interaction, vol. 1, M. G. Helander, T. K. Landauer, and P. V. Prabhu, Eds. Amsterdam: Elsevier, 1997, pp. 1003-1041.

[23] Gaver, W. W., "Sound Support for Collaboration", In proceedings of Second European Conference on Computer-Supported Cooperative Work, Amsterdam, pp. 293-308, 1991.

[24] Gaver, W. W., Smith, R. B., and O'Shea, T., "Effective Sounds in Complex Systems: The ARKOLA Simulation", In proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Reaching through Technology New Orleans, Louisiana, USA, pp. 85-90, 1991.

[25] Hauptmann, A. G. and Rudnicky, A. I., "A Comparison of Speech and Typed Input", In proceedings of the Speech and Natural Language Workshop, San Mateo, California, pp. 219-224, 1990.

[26] Jung, J. H., Looney, C. A., and Valacich, J. S., "Fine-Tuning the Human-Computer Interface: Verbal versus Keyboard Input in an Idea Generation Context", In proceedings of the 40th Hawaii International Conference on System Sciences (HICSS'07), pp. 27c, 2007.

[27] Kirby, M. A. R., "Systems Design: Ensuring the Right Users are Involved", In proceedings of the 8th International Conferrence on Human-Computer Interaction HCI '99, Munich, Germany, pp. 169-170, 1999.

[28] Lahtinen, S. and Peltonen, J., "Enhancing Usability of UML CASE-Tools with Speech Recognition", In proceedings of IEEE Symposium on Human Centric Computing Languages and Environments, pp. 227-235, 2003.

[29] Leplatre, G. and Brewster, S. A., "Designing Non-Speech Sounds to Support Navigation in Mobile Phone Menus", In 6th International Conference on Auditory Display (ICAD), Atlanta, Georgia, USA, pp. 190-199, 2000.

[30] Lindberg, T. and Nasanen, R., "The effect of icon spacing and size on the speed of icon processing in the human visual system", Displays, vol. 24, pp. 111-120, 2003.

[31] Lumsden, J., Brewster, S., Crease, M., and Gray, P. D., "Guidelines for Audio-Enhancement of Graphical User Interface Widgets", In proceedings of BCS HCI'2002, London, 2002.

[32] Manaris, B., "Natural Language Processing: A Human-Computer Interaction Perspective", Advances in Computers, vol. 47, pp. 1-66, 1996.

[33] Manaris, B., "SUITEKeys: A Speech Understanding Interface for the Motor-Control Challenged", In ACM SIGACCESS Conference on Assistive Technologies Marina del Rey, California, USA, pp. 108 - 115, 1998.

[34] Manaris, B., McCauley, R., and MacGyvers, V., "An Intelligent Interface for Keyboard and Mouse Control--Providing Full Access to PC Functionality via Speech", In proceedings of the Fourteenth International Florida Artificial Intelligence Research Society Conference, pp. 182-188, 2001.

[35] Marentakis, G. and Brewster, S. A., "A Comparison of Feedback Cues for Enhancing Pointing Efficiency in Interaction with Spatial Audio ", In proceedings of the 7th international conference on Human computer interaction with mobile devices & services, Salzburg, Austria, pp. 55 - 62, 2005.

[36] Marentakis, G. and brewster, S. A., "Effects of Feedback, Mobility and Index of Difficulty on Deictic Spatial Audio Target Acquisition in the Horizontal Plane", In proceedings of the SIGCHI conference on Human Factors in computing systems, Montréal, Québec, Canada, pp. 359 - 368, 2006.

[37] Mellor, B. A., Baber, C., and Tunley, C., "Evaluating Automatic Speech Recognition as a Component of a Multi-Input Human-Computer Interface", In proceedings of the International Conference on Spoken Language, 1996.

[38] Nicol, C., Brewster, S. A., and Gray, P. D., "A System for Mainpulating Auditory Interfaces Using Timbre Spaces", In proceedings of CADUI 2004, Madeira, Portugal, pp. 366-379, 2004.

[39] Oakley, I., McGee, M. R., Brewster, S., and Gray, P. D., "Putting the feel in look and feel", In ACM CHI 2000 (The Hague, NL), pp. 415-422, 2000.

[40] Oviatt, S. and Cohen, P., "Designing the User Interface for Multimodal Speech and Pen-Based Gesture Applications: State-of-the-Art Systems and Future Research Directions", Human-Computer Interaction, vol. 15, pp. 263-322, 2000.

[41] Oviatt, S., Coulston, R., and Lunsford, R., "When do we interact multimodally?: Cognitive Load and Multimodal Communication Patterns", In proceedings of the 6th international conference on Multimodal interfaces State College, PA, USA, pp. 129 - 136, 2004.

[42] Oviatt, S. L., "Taming Recognition with Multimodal Interface", Communications of the ACM, vol. 43, pp. 45-51, 2000.

[43] Rigas, D. and Memery, D., "Using Non-Speech Sound to Communicate Information in User Interfaces", In Applied Informatics 2000, Innsbruck, Australia, pp. 357-362, 2000.

[44] Rigas, D., Memery, D., and Yu, H., "Experiments in Using Structured Musical Sound, Synthesised Speech and Environmental Stimuli to Communicate Information: Is there a Case for Integration and Synergy?" In proceedings of International Symposium on Intelligent Multimedia: Video and Speech Processing, Kowloon Shangri-La, Hong Kong, pp. 465-468, 2001.

[45] Rigas, D., Yu, H., Klearhou, K., and Mistry, S., "Designing Information Systems with Audio-Visual Synergy: Empirical Results of Browsing E-Mail Data", In Panhellenic Conference on Human-Computer Interaction: Advances on Human-Computer Interaction, Patras, Greece, pp. 960-7620-18, 2001.

[46] Rigas, D., Yu, H., and Memery, D., "Experiments Using Speech, Non-Speech Sound and Stereophony as Communication Metaphors in Information Systems", In proceeding of the 27th Euromicro Conference, Warsaw, Poland, pp. 383-390, 2001.

[47] Rigas, D., Yu, H., Memery, D., and Howden, D., "Combining Speech with Sound to Communicate Information in a Multimedia Stock Control System", In 9th International Conference on Human-Computer Interaction: Usability Evaluation and Interface Design, New Orleans, Luisiana, USA, pp. 1262-1266, 2001.

[48] Rigas, D. I., "Audiotest: Utilising Audio to Communicate Information in Program Debugging", In proceedings of the 8th International Conferrence on Human-Computer Interaction HCI '99, Munich, Germany, pp. 1293-1297, 1999.

[49] Rigas, D. I., Hopwood, D., and Memery, D., "Communicating spatial information via a multimedia-auditory interface", In EUROMICRO Conference, 1999. Proceedings. 25th, pp. 398-405 vol.2, 1999.

[50] Rigas, D. I. and Alty, J. L., "Using Rising Pitch as a Communication Metaphor: An Empirical Investigation", In Euromicro Conference, 2000. Proceedings of the 26th, pp. 322-331 vol.2, 2000.

[51] Rigas, D. I., Memery, D., Hopwood, D., and Rodrigues, M. A., "Using Non-Speech Sound to Communicate Information in User Interfaces", In Applied Informatics 2000, Innsbruck, Austria, pp. 357-362, 2000.

[52] Rigas, D. I., Memery, D., Hopwood, D., and Rodrigues, M. A., "Empirically derived design issues in auditory information processing for mobile telephony", In Information Technology: Coding and Computing, 2000. Proceedings. International Conference on, pp. 462-469, 2000.

[53] Rigas, D. I. and Memery, D., "Utilising Audio-Visual Stimuli in Interactive Information Systems: A Two Domain Investigation on Auditory Metaphors", In International Conference on Information Technology: Coding and Computing, pp. 190-195, 2002.

[54] Rigas, D. I. and Atly, J. L., "The Rising Pitch Metaphor: An Empirical Study", International Journal of Human-Computer Studies, vol. 62, pp. 1-20, 2005.

[55] Seagull, F. J., Wickens, C. D., and Loeb, R. G., "What is Less More? Attention Workload in Auditory, Visual, and Redundant Patient-Monitoring Conditions", In proceedings of 45th Annual Meeting of the Human Factors and Ergonomics Society, Santa Monica, Canada, 2001.

[56] Vargas, M. L. and Anderson, S., "Combining Speech and Earcons to Assist Menu Navigation", 2003.

[57] Walker, B. N. and Kramer, G., "Ecological Psychoacoustics and Auditory Displays: Hearing, Grouping, and Meaning Making", Ecological Psychoacoustics, pp. 150-175, 2004.

[58] Zhou, L., Feng, J., Sears, A., and Shi, Y., "Applying the Naïve Bayes Classifier to Assist Users in Detecting Speech Recognition Errors", In proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05), pp. 1-9, 2005.