Conversational Gestures For Direct Manipulation On The Audio Desktop

T. V. Raman
Advanced Technology Group
Adobe Systems
E-mail: ⟨raman@adobe.com⟩
WWW: **http://cs.cornell.edu/home/raman**

Abstract

We describe the speech-enabling approach to building auditory interfaces that treat speech as a first-class modality. The process of designing effective auditory interfaces is decomposed into identifying the atomic actions that make up the user interaction and the conversational gestures that enable these actions. The auditory interface is then synthesized by mapping these conversational gestures to appropriate primitives in the auditory environment.

We illustrate this process with a concrete example by developing an auditory interface to the visually intensive task of playing tetris. Playing Tetris is a fun activity[1] that has many of the same demands as day-to-day activities on the electronic desktop. Speech-enabling Tetris thus not only provides a fun way to exercise ones geometric reasoning abilities —it provides useful lessons in speech-enabling common-place computing tasks.

[1]This paper was seriously delayed because the author was too busy playing the game.

ASSETS 98 Marina del Rey CA USA

1 Introduction

The phrase *desktop* no longer conjures up the image of a polished high-quality wooden surface. The pervasiveness of computing in the workplace during the last decade has led to the concept of a virtual *electronic desktop* —a logical workspace made up of the documents one works with and the applications used to operate on these documents. Progressive innovations in the Graphical User Interface (GUI) have helped strengthen this metaphor —today, the typical *desktop* enables the user to organize the *tools of his trade* by dragging and dropping graphical icons into a visual two dimensional workspace represented on the computer monitor. Given this tight association between visual interaction and today's electronic desktop, the phrase *audio* desktop is likely to raise a few eyebrows (Or should it be ear lobes)!

This paper focuses on specific aspects of auditory interaction with a view to enabling an audio desktop. The audio desktop is defined in detail in Chapter 3 of [Ram97a]. Using the speech-enabling approach first introduced in [Ram96a, Ram96b, Ram97b], we demonstrate how the functionality of the electronic desktop can be exposed through an auditory interface. The attempt is not to *speak* the visual desktop; rather, we identify the key user-level functionality enabled by the modern electronic desktop and briefly describe how this can be

translated to an auditory environment.
This paper specifically illustrates the auditory analogue to gestures available on the visual desktop by describing an auditory interface to the popular game of Tetris. For a detailed overview of a full implementation of an auditory desktop, see Chapter 4 of [Ram97a]. Though speech-enabling a game like Tetris might seem a somewhat light-hearted (and perhaps even pointless) activity, there are important lessons to be learned from speech-enabling such a visually intensive task. Many of the demands placed on the user by a game like Tetris are closely paralleled by the functional abilities demanded by today's computer interfaces. Speech-enabling Tetris thus provides a fun activity on the surface while exposing deeper research ideas that have wide-ranging applicability in the general design of auditory interfaces to tomorrow's information systems.
In visual interaction, the user actively browses different portions of a *relatively* static two dimensional display to locate and manipulate objects of interest. Such manipulations are aided by hand-eye coordination in visually intensive tasks like playing Tetris. Contrast this with auditory displays that are characterized by the temporal nature of aural interaction; here, the display —a one-dimensional stream of auditory output— *scrolls* continuously past a passive listener. This disparity between aural and visual interaction influences the organizational paradigms that are effective in auditory interaction. The purpose of this paper is to systematically investigate the design of an effective audio interaction to a visually intensive task like playing Tetris. The steps in evolving such an interface can be enumerated as:

- Identify user functionality enabled by the visual interface,
- Exploit features of auditory displays to enable equivalent functionality and
- Evolve navigational and organizational paradigms for aural interaction that compensate for the temporal, one-dimensional nature of audio by exploiting other features of aural interaction.

2 Conversational Gestures

User interface is a means to enabling man-machine communication. This man-machine *dialogue* takes place by means of a set of simple conversational gestures designed to overcome the impedance mismatch in the abilities of man and machine. These gestures are realized in today's Graphical User Interfaces (GUIs) by user interface widgets such as list boxes and scroll bars.

Natural Language			
Edit widgets	Message widgets		
Answering Yes Or No			
Toggles	Check boxes		
Select From Set			
Radio groups	List boxes		
Select From Range			
Sliders	Scroll Bars		
Traversing Complex Structures			
Previous	Next	Parent	Child
Left	Right	Up	Down
First	Last	Root	Exit

Figure 1: Conversational gestures —the building blocks for dialogues. User interface design tries to bridge the impedance mismatch in man-machine communication by inventing a basic set of conversational gestures that can be effectively generated and interpreted by both man and machine.

We first enumerate the basic conversational gestures that constitute today's user interfaces in figure 1. Separating conversational gestures *e.g., select an element from a list* from the

modality-specific realization of that gesture —a list box in the case of the GUI— (see Figure 1 on the preceding page) is the first step in evolving speech-centric man-machine dialogues.

3 The Game Of Tetris

This section briefly describes the game of Tetris and enumerates the conversational gestures involved in playing the game. These gestures are introduced with respect to the familiar visual interface; later sections translate these to appropriate gestures in an auditory interface. The game involves forming rows by arranging interlocking shapes. When complete these rows disappear from the board. Tetris shapes are the seven possible arrangements of four square tiles —see Figure 2. The shapes drop from the top of the screen, and the user has to move and rotate the shape before dropping it to fit in with those at the bottom of the playing area. In this paper, we consider an instance of the game where the playing area is ten columns wide and twenty rows high.

Playing Tetris involves geometric reasoning to decide where best to fit the current tile. In the visual interface, the user can use the two-dimensional nature of the visual display backed up by hand-eye coordination to line up the current shape with the available openings on the bottom row. The conversational gestures involved are:

Indicate Current Shape The current shape is indicated to the user by dropping it from the top of the playing area.

Choose Location The user examines the available openings on the bottom row to mentally construct a set of available positions that the shape can be placed in.

Choose Orientation The user selects a valid orientation for the shape and its chosen location.

Fit Shape The user fits this shape at the chosen location and is given the next shape.

Update State If fitting this shape completed a row, the user is cued appropriately. The playing area is redrawn to indicate the available openings for the next shape.

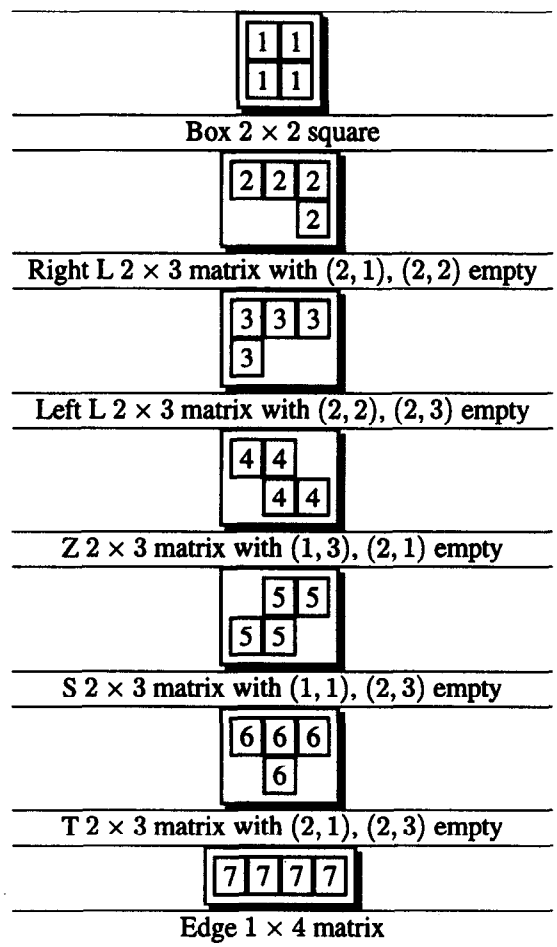

Figure 2: The seven shapes of Tetris.

4 Direct Manipulation In A Visual Environment

Playing Tetris exercises one of the basic functionalities of the electronic desktop,

namely, the user's ability to directly manipulate objects in the interface. Here, the user expresses actions by *selecting* and *moving* the shape with a pointing device or appropriate keyboard events. The continuous visual feedback loop that is a direct consequence of hand-eye coordination enables the user to line up the shape with the available openings on the bottom of the playing area. Using different colors for the various shapes helps the user quickly identify each new shape as it arrives. The eye's ability to quickly scan different portions of the two-dimensional display helps the user identify possible locations for the current shape.

5 Direct Manipulation In An Auditory Environment

The temporal one-dimensional nature of aural interaction can be a significant hurdle in attempting a visually intensive task such as playing Tetris. We compensate for these shortcomings by enabling appropriate gestures in the auditory interface that permit the user to obtain the necessary information and express appropriate actions in a timely and effective manner. See Table 1 on page 6 for a full list of commands provided in the auditory interface.

The design of the auditory interface to Tetris is predicated by the following:

- The geometric reasoning required for fitting interlocking shapes can be carried out mentally once the user has been given sufficient information.

- It is possible to translate actions predicated by visual geometric reasoning such as "move a little to the left" to functionally precise actions such as "move two steps left" given sufficient information.

5.1 Conveying The Shapes

The seven shapes of Tetris are shown in Figure 2 on the page before. Each shape is given a mnemonic name based on its visual appearance. These mnemonics are used when announcing the current shape in the auditory interface. Thus the user hears spoken utterances of the form

> Left Elbow at rotation 0 next is Right Elbow.

We use the digits 1–7 as *functional colors* in place of physical colors such as red and blue. This helps the listener recall what shapes were fitted when examining the state of the game. Each shape in figure Figure 2 on the preceding page is accompanied by a detailed verbal description to make the information readily accessible when reading this paper in alternative formats.

5.2 Expressing Actions

The auditory interface enhances the available gestures by providing keyboard commands for moving the current shape to a given absolute position. This is the single most important enhancement that the auditory interaction needs over visual interaction. Unlike in visual interaction, a user of the auditory interface does not have the continuous visual feedback loop that allows the current shape to be lined up with the available openings. In the case of the auditory interface, the listener needs to mentally track these openings in order to play the game fluently. Having to then line up the current shape using only relative translations becomes an undue mental burden. The ability to position the shape with absolute coordinates, *e.g.*, "move to column 3", compensates for this shortcoming and allows the user to play the game effectively.

5.3 Providing Feedback

Providing prompt and immediate feedback is essential in providing effective interaction. The auditory interface indicates the dropping of each shape to the bottom with an auditory icon. As the user drops each shape, the system produces a distinctive click; when the piece drops to form a complete row, this click is replaced by a short chime. Use of such auditory icons (each cue is about 0.5 seconds long) is extremely effective in designing fluent aural interaction.

5.4 Communicating state of the game

The two-dimensional display allows a user of the visual interface to implicitly query different aspects of the state of the game such as

- What does the top row look like?
- What does the bottom row look like?
- How high is the stack of shapes?
- How well am I currently doing?

with simple eye movements. In the auditory interface, we make these actions explicit by providing keyboard actions that speak the response to these queries —see Table 1 on the next page for a complete list.

Key	Action
Relative Motion	
h	Move left
l	Move right
j	Rotate counter-clockwise
k	Rotate clockwise
Absolute Motion	
a Move to left edge e Move to right edge Digit Move to column ⟨Digit⟩	
Examine State	
b	Bottom row
t	Top row
c	Current row
m	Current row
r	Row number
.	Current shape
,	Next shape
RET	Score

Table 1: Complete list of commands in the auditory interface to Tetris

6 Conclusion

By systematically enumerating the atomic actions that happen in a visually intensive activity like playing tetris, and mapping these to a basic set of conversational gestures is the first step in speech-enabling this game. Mapping basic conversational gestures to appropriate events in an auditory interface leads to a speech-enabled version of Tetris. The process of evolving this interface has important lessons for designing auditory interfaces to day-to-day tasks on the electronic desktop. Primary among these are:

- Enable user to express intent precisely.
- Provide sufficient feedback to enable the user to maintain a mental model that is synchronous with the state of the computing system.
- Use auditory cues (see [RK92, SMG90, BGP93]) and audio formatted output (see [Ram94, RG94, Gib96, Hay96]) to increase the band-width of aural communication.

References

[BGP93] Meera M. Blattner, Ephraim P. Glinert, and Albert L. Papp.
Sonic Enhancements for 2-D Graphic Displays, and Auditory Displays.
To be published by Addison-Wesley in the Santa Fe Institute Series. IEEE, 1993.

[Gib96] Wayte Gibbs.
Envisioning speech.
Scientific American, September 1996.

[Hay96] Brian Hayes.
Speaking of mathematics.
American Scientist, 84(2), March–April 1996.

[Ram94] T. V. Raman.
Audio System for Technical Readings.
PhD thesis, Cornell University, May 1994.
URL **http://cs.cornell.edu/home/raman.**

[Ram96a] T. V. Raman.
Emacspeak —direct speech access.
Proc. of The Second Annual ACM Conference on Assistive Technologies (ASSETS '96), Apr 1996.

[Ram96b] T.Ṽ. Raman.
Emacspeak –a speech interface.
Proceedings of CHI96, April 1996.

[Ram97a] T. V. Raman.
Auditory User Interfaces –Toward The Speaking Computer.
Kluwer Academic Publishers, August 1997.

[Ram97b] T. V. Raman.
Net surfing without a monitor.
Scientific American, March 1997.

[RG94] T.V. Raman and David Gries.
Interactive audio documents.
Proc. 1st Annual ACM/SIGCAPH Conf. on Assistive Technology, Nov 1994.

[RK92] T. V. Raman and M. S. Krishnamoorthy.
Congrats: A system for converting graphics to sound.
Proceedings of IEEE on Johns Hopkins National Search for Computing Applications to Assist Persons with Disabilities, pages 170–172, February 1992.

[SMG90] D. A. Sumikawa, Blattner M. M., and R. M. Greenberg.

Earcons and icons: Their structure and common design principles. *Visual Programming Environments*, 1990.