Music and speech in auditory interfaces: When is one mode more appropriate than another?

Research paper for the ICAD05 workshop "Combining Speech and Sound in the User Interface"

James L Alty

School of Computing Science,
Middlesex University
London L17 8HR
UK.
j.l.alty@mdx.ac.uk

Dimitrios Rigas

School of Informatics.
Bradford University
Bradford
UK
D.Rigas@bradford.ac.uk

Paul Vickers

School of Computing, Engineering and Information Science
Northumbria University
Newcastle upon Tyne, NE2 1XE
UK
P.Vickers@northumbria.ac.uk

ABSTRACT

A number of experiments, which have been carried out using non-speech auditory interfaces, are reviewed and the advantages and disadvantages of each are discussed. The possible advantages of using non-speech audio media such as music are discussed – richness of the representations possible, the aesthetic appeal, and the possibilities of such interfaces being able to handle abstraction and consistency across the interface.

1. INTRODUCTION

Continuous speech has been used from the early days in human-computer interface design. For example, it was used extensively to provide support for blind or visually impaired users (its use in Screen Readers), and if the information to be communicated is in textual form, speech is almost certainly a most appropriate way of communicating it via the auditory channel. However, there are many other types of information besides text which need to be communicated and attempts have also been made to communicate these utilising simple musical sequences, or natural auditory sounds but, for the most part, the only information communicated using these media has been the occurrence of simple events.

There are good reasons why interfaces employing music have not been fully explored. In contrast to speech, it is difficult to use music to communicate quantitative information. In addition there are likely to be both significant cultural issues and concerns about the degree of musical ability required from the user population to make sense of a musically based interface. Furthermore, in the early days of computing, the creation of realistic musical sounds was not possible. In the 1970s, however, the technology significantly advanced with the development of the MIDI Interface [1]. Even a relatively inexpensive Personal Computer will now have a good sound card and easy access to a set of realistic musical sounds. Furthermore the development of sequencers such as Sibelius [2] and Cubase [3] have enabled designers to create very realistic orchestrations.

In some of our early work with blind and partially sighted users, we repeatedly came across comments from them criticising speech based interfaces, particularly if the information conveyed was complex and not just concerned with the communication of text. They often commented that speech-based information was tiring to listen to, particularly if it was trying to communicate visual information. Furthermore, we have argued that music offers a powerful medium for communication [4] and have looked for ways to use its structures and organisational principles to better communicate program information. The key issue is how to map domain entities to musical structures.

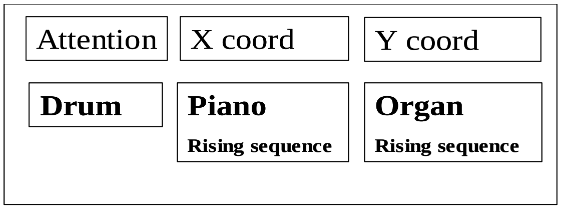
As a result, we deliberately attempted in a series of experiments to create interfaces which relied entirely on music, and which had no speech component at all. Three major experiments were carried out:

- The construction of a system for communicating graphical information using music alone (called AUDIOGRAPH)
- The construction of a system for assisting novice programmers with the debugging task in PASCAL using musical motifs (called CAITLIN)
- The auralisation of algorithms using music alone

2. THE AUDIOGRAPH SYSTEM

The initial research work created a musically based system for communicating diagrammatic information for the blind called AUDIOGRAPH [6], [7], [8]. In this system, the user is presented aurally with a 40x40 grid that can contain graphical objects such as squares, circles, rectangles and lines. All communication is through musical sequences alone.

A coordinate location (x, y) within the 40 x 40 graphical grid is communicated using a pair of sequences of rising notes (range: 1 to 40). The first sequence of notes (using a piano) always communicated the horizontal (x) coordinate and the second sequence (using an organ) communicated the vertical (y) coordinate. Both sequences always started from the same root note E_2 and advanced up the Chromatic Scale (i.e. the black and white notes on the piano) to the same final note A_5 (A flat). A long sequence of notes communicated high coordinate values and vice-versa. For example, the coordinate position of an object in the space is communicated as in Figure 1.



Attention	X coord	Y coord
Drum	**Piano** **Rising sequence**	**Organ** **Rising sequence**

Figure 1. *Musical presentation of a coordinate*

The structure of the various graphical shapes was communicated by following the outline of the shapes and playing the sequence of notes that described the coordinates of the shape. Figure 2 illustrates some examples.

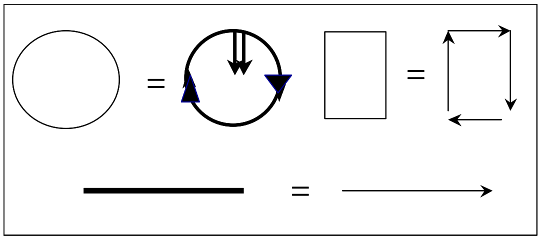


Figure 2. *Communication of the Shapes*

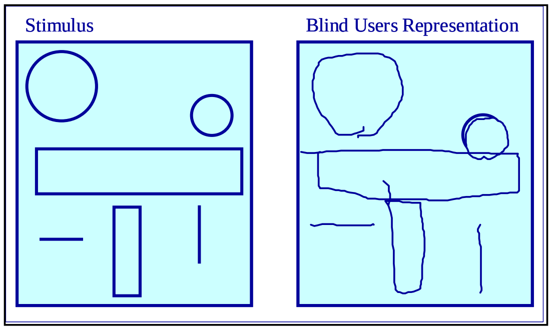


Figure 3. *A Blind User's interpretation of a Stimulus.*

Although the shapes took a little time to play, it was remarkable how quickly the users were able to recognise the different shapes.

The system could be used both for communicating a set of shapes and the user could also alter them (move, expand, contact, add a new shape etc.). Without using any text at all users were able to understand the basic shapes contained in the grid. For example, Figure 3 shows the visual view of a presentation (communicated aurally) and a blind user's attempt at drawing what he heard.

Although the positioning is not exact, it is quite close. The lines are rather shaky because the users have to draw what they hear on a raised grid and this often deflects the pencil.

3. THE CAITLIN SYSTEM

A second example of an experiment we carried out using only musical output is the CAITLIN musical program auralization system [9], [10], [11] [12]. This system demonstrated that a musical auralization framework for communicating run-time behaviour of PASCAL programs was successful in assisting users with bug location tasks. Rendering the workings of a computer programming poses some interesting problems. Program events happen in the time domain whilst visual techniques give us better descriptions of spatial relations and structural details. Sound on the other hand can present us with a temporal view of the software operations (as a wave-form plot does for a sound wave). Thus an auditory approach might yield some interesting alternative views of the debugging process. In this particular experiment the users had normal sight, but debugged the programs using music alone.

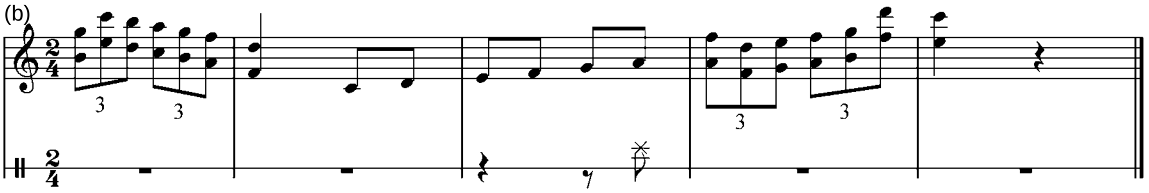
The basic concept of the CAITLIN system involved taking the source code of a PASCAL program and adding a set of MIDI commands (via a pre-compiler) to the program structures so that when the program executed it also created a sequence of musical sounds, otherwise, the program code is left unchanged. Pauses had to be added to prevent the music from playing too fast, though we allowed users to vary the speed of presentation. The contructs which are mapped to sound representations are of two basic types – iteration and selection. Selection involves the constructs of IF and CASE (including IF..THEN, IF..THEN..ELSE, CASE, and CASE..ELSE). Iteration involves the structures of FOR..TO, FOR..DOWN TO, WHILE and REPEAT.

Each structure was represented by an introductory motif which was repeated in reverse order at the end (rather like the IF..FI constructions in pseudocode). So the IF contruct was introduced by a short sound to alert the user, and then the 1-bar motif representing IF was played. An IF statement evaluates a Boolean expression to either

(a)

```
FOR counter := 1 TO 6 DO
   counter := counter + 1 ;
```

The Pascal code (a) results in the auralization (b). Bar 1 and beat 1 of bar 2 is the tune denoting entry to the loop. The notes remainder of bar 2 and bar 3 represent six iterations of the loop, the final iteration being supplemented by a sleigh bell sound. Bars 4 and 5 denote exit from the loop.



(c)

```
a : = 10 ;
CASE a OF
   '1' : Writeln ('Found 1') ;
   '2' : Writeln ('Found 2') ;
   '3' : Writeln ('Found 3') ;
   '4' : Writeln ('Found 4') ;
   ELSE  Writeln ('Not found)
END ;
```

The Pascal code (c) results in the auralization (d). Bar 1 and beat 1 of bar 2 is the tune denoting entry to the CASE construct. A cowbell sound is played as each case instance is tested (bars 2, 3, and 4). The tune in bars4 to 6 signify exit from the construct. In this example, no match was found for the case selector, so the construct exits in a minor key. In the CAITLIN system the major mode was used to denote Boolean true and minor for Boolean false.

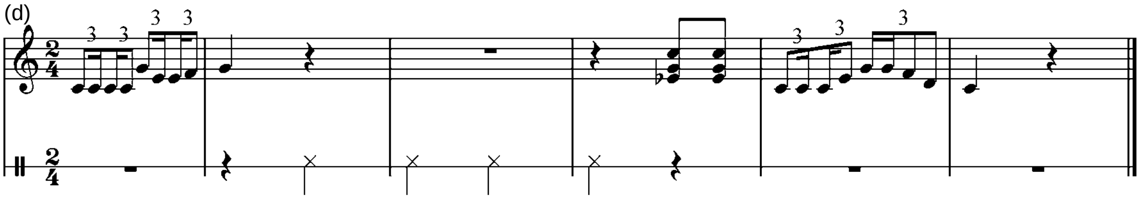


Figure 4. *Two Examples of Constructs used in the CAITLIN System*

TRUE or FALSE. A major or minor chord was sounded to show a successful conclusion or failure of that Boolean expression. Finally, a reversed version of the motif was played to indicate the conclusion of the construct. A similar approach was used for all Selection constructs though the musical motifs were similar but distinct. In the case of the Iteration constructs, the musical representation has to show the progression through the loops. This was achieved by holding a drone note until the Boolean expression was evaluated as TRUE or FALSE (again showed by a major or minor chord). Two typical examples of the musical motifs used in the CAITLIN system are shown in Figure 4 - a FOR loop and a CASE statement (taken from [12].

The design of the various motifs allowed constructs to be nested upto three in depth. Thus a user could hear the start of a FOR loop, then followed an IF statement which in turn contained a CASE statement (or any other combination).

The various papers [9], [10], and [11] provide an in-depth analysis of the results obtained. Some constructs were easier to identify than others and this may have been influenced by the choice of musical motif used. The work on the CAITLIN system concluded with a debugging study using the constructs and the motifs previously described. The bugs introduced all affected program flow (since many other aspects of program execution were not musically represented).

The two major classes of bugs investigated were:

a) Ill-formed simple Boolean expressions directly causing perturbation in the program flow
b) Incorrect assignments that manifest themselves indirectly through program flow.

Half the subjects were presented with the auralisation representations and the other half with visual representations. Speech was not used in any part of the experiment. Other factors measured included the "annoyance factor" of using music and the workload as measured by the NASA task load index [13]. The information about program structure and that the auralisations could be interpreted at different levels. Also, for some types of bug, musical information could assist novice programmers in locating pre-planted bugs in short PASCAL programs.

4. ALGORITHM AURALISATION

Alty [5] showed that algorithms (such as the bubble sort and minimum-path) can have information about their run-time behaviour communicated successfully through musical mappings. In the Bubble Sort Example, the state of the list is mapped to a rising pitch on a Clarinet, a rising note on the Harps used to indicate which pair of numbers in the list are being considered, a flurry of trumpets indicates a swap, and a musical cadence signifies the completion of the process. Stereophony was also used to assist in distinguishing between the parallel set of musical indicators.

The demonstration worked well, and most subjects (whether musically trained or not) quickly understood what was going on in the algorithm.

The results suggest that, provided precise numerical relationships are not being communicated, music can transfer information successfully. Furthermore, in these early experiments comparisons were made between the performance of musically trained and musically untrained subjects and no significant differences were observed.

5. SPEECH OR MUSIC?

Speech initially would appear to have many advantages over music and other sounds. Firstly, it is a normal form of communication between human beings. It can accurately communicate exact values for numerical values and precise meaning for text. Music has neither of these advantages. However, our experiments have shown that the situation is not as clear-cut as these considerations might imply.

Firstly, there are many situations where communication need not be precise or exact. For example, when using the AUDIOGRAPH system, users were only able to determine shape position to about 90% accuracy and although they were always able to identify the graphical shape (square, circle, rectangle, line), they would often only determine the approximate size. However, this is not a problem for many graphical actions. For example, if a sighted user was asked to look at a graphical diagram and then draw it separately, they would also make similar inaccuracies in the reproduced diagram.

When studying a map, say, of a city centre, the important thing is to correctly determine the relative position of shops or roads rather than know precisely where they are in relation to each other.

Although music cannot be used to communicate precise numerical information, our experiments have shown that it can be used to give quite good relative information. For example, Figure 5 shows how users responded in AUDIOGRAPH to the presented coordinates (with values from 1 to 40).

Although the inaccuracies can clearly be seen, from a relative point of view the approach works quite well. We therefore concluded that music can be used to communicate numerical information provided relative accuracy is the main criterion for success.

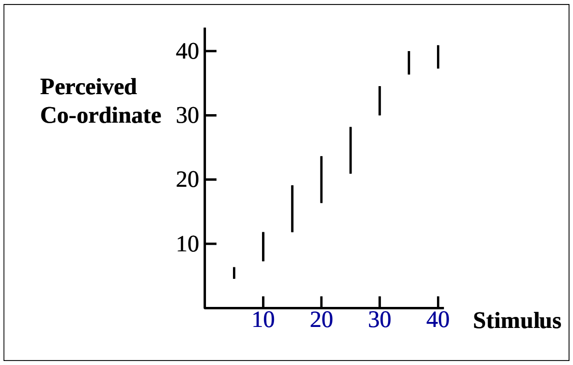


Figure 5. *User response to musical presentation of coordinates in AUDIOGRAPH.*

But why bother at all with music? What advantages does it offer over the use of speech? We have found that there are a number of potential advantages:

a) Speech used for any length of time can be very tiring and irritating unless the information being communicated is meaningful text. This is not just a problem with the creation of synthetic speech. Auditory descriptions can be difficult to follow if complex.
b) Music is intrinsically enjoyable to listen to. Users often find listening to musical information more relaxing.
c) The different possibilities offered through musical design (rhythm, harmony, pitch, timbre) provide a rich set of contrasting information sources which often can

highlight really important information in a way that speech cannot.

d) Speech does not offer the possibilities of abstraction. For example, it is not possible to "zoom in" and "zoom out" to get a contrasting view of what is being communicated.

5.1 Irritation or Annoyance.

It does seem to be true that if aesthetic considerations are taken into account, auditory displays are usually much easier to listen to and to comprehend. For example, Mayer-Kress and workers [14] have mapped chaotic attractor functions to musical structures in which the functions' similar but never-the-same regions could be clearly heard. The resultant music was able to be appreciated in its own right without needing to know how it was produced. Another example is Quinn's *Seismic Sonata* (2000) [15] which uses the aesthetics of musical form to sonify data from the 1994 Northridge, California earthquake. However, acceptance does depend critically on the nature of the musical phrases being communicated. For example, in the CAITLIN system, the perceived workload using the auralisations was significantly greater.

The other possible effect of musical auralisations is annoyance. In the CAITLIN experiments subjects were asked to rate the degree of annoyance caused by the auralisations. About half the subjects found the auralisations moderately annoying, but the other half found them to be acceptable.

5.2 Enjoyableness of Music

What surprised us in the presentation of the Auralised algorithms was the degree of acceptance by the audiences they were played to. In many cases the audience actually applauded at the conclusion of the demonstration! Blind users also commented favourably on the auralisations used in the AUDIOGRAPH system. This implied to us that music offered considerable advantages over non-musical audio. In one sense this result is not totally surprising. Music forms an important part of most peoples lives. So, provided the musical representations follow conventional harmonic and rhythmical rules, it might be expected that they would be found to be enjoyable.

This, however, identifies an important design constraint in the development of musical interfaces – a musically trained designer must be involved. In our case it was fortunate that one of the authors was musically trained both in performance and in composition, and another author was a competent musician.

5.3 Richness of the Representations

Music is the most highly developed of the auditory media. It is therefore very rich in usable representations. The possible variables that can be used include rhythm, pitch and timbre, and combinations of this such as harmony (chords), polyphony and part writing.

Pitch alone provides a useful way of representing numbers. A typical musical scale can accommodate 70 to 80 different numerical values. In the past the use of pitch has been criticised because of what is called the "octave effect". Users can have difficulty in distinguishing between the same note played in a different octave. However, the experiments that suggested this effect were usually carried out using pure sine waves as stimuli, and it is not surprising that the effect was observed in this situation. In real music, composers use instruments whose harmonic content varies over the scale. Thus, for an instrument such as a clarinet, the note C_3 will sound quite distinctive compared with the same note from different octaves such as the notes C_4 or C_5.

One powerful variable in music is timbre (or the sound of an instrument). For example, a flute will sound very different from a piano. However, some timbres are quite close – for example a cello and a violin. How can a designer be sure that users (particularly non-musically trained users) will be able to distinguish between presented timbres?

This was a problem that we examined when designing the AUDIOGRAPH interface. We played different timbres to users and asked them to identify them. There was confusion over a number of timbres but when we separated them into the musical families used by composers (e.g. wind, brass, timpani, strings etc.) the confusion was considerably reduced. Figure 6 illustrates the results over the main musical families.

	Piano	**Wind**	**Timpani**	**Brass**	**Strings**	**Organ**
Piano	**65**	**0**	**0**	**2**	**3**	**4**
Wind	**4**	**59**	**0**	**5**	**13**	**15**
Timpani	**0**	**0**	**15**	**0**	**0**	**0**
Brass	**8**	**7**	**0**	**18**	**11**	**9**
Strings	**2**	**9**	**0**	**5**	**14**	**2**
Organ	**1**	**4**	**0**	**3**	**5**	**15**

Figure 6. *Distinguishing Timbres between families*

Notice how distinct timpani (drums) and piano are. Brass on the other hand can be mistaken for other timbres. One interesting point about these results is that they were measured using the output of a standard audio card. If a more sophisticated card had been used (or sampled sounds of real instruments) the distinctions would have been clearer.

Another interesting representation in music that can be explored is polyphony. This is the creation of musical lines running in parallel (one classic example being "Frere Jaques" sung by four groups with each line displaced in time). Human beings are quite adept at hearing different lines simultaneously and this can be used effectively when trying to distinguish between parallel sequences of events.

5.4 Degree of Abstractness

Music's ability to represent the same set of events at different levels of abstraction is probably one of the most compelling reasons for using it in preference to speech. With visual interfaces, different levels of abstraction can be shown using zooming. For example, if a map is being examined, it can be viewed at an individual street level or at a town level. In music we have found that the equivalent technique is using speed of presentation. For example, if an audiolisation of a Bubble Sort is presented at a faster speed, some of the detailed information is filtered out and the user hears a higher level representation. We suspect that this would also be true in the Debugging output.

Another useful aspect of using music is that a common representation can be used across an application. For example, in the AUDIOGRAPH, the same musical representation is used for displaying where the cursor is, the actual coordinate location of a shape, and the tracing of the shape of an object. Another common musical representation is used for representing the commands Expand, Contract and Delete. This resulted in the commands being learned quickly and users could even make intelligent guesses as to the nature of a command even if they had not heard it before.

6. CONCLUSIONS

Provided the limitations of non-speech representations are appreciated (i.e. representation of exact numerical values), they can be used successfully in a wide variety of applications. They also have advantages over speech in that they can be aesthetically more pleasing, can be used to provide a common representation system across applications, and can provides the ability to present information at different levels of abstraction. Many of the objections to non-speech media are not as significant as was once thought. Although cultural differences may occasionally be significant, the commonality across musical cultures is greater than the differences. For example, most music systems use the same set of notes or subsets of them (e.g. the pentatonic scale). Furthermore, it appears that the musical ability of the average person is well able to understand the interfaces which we created.

There are still issues with respect to workload and irritation which need to be addressed, but the work done so far seems to indicate that non-speech auditory information could be useful both for normally sighted users as well as those visually challenged.

What has not been done in any of the above work is a direct comparison of user performance when presented with either interfaces using speech or with interfaces utilizing non-speech audio. What we have shown is that non-speech audio can be used successfully to communicate information in certain tasks and that the information communicated is much more than the occurrence of simple events.

7. REFERENCES

[1] MIDI-1.0 Specification, *Roland Corp. US.*, 7200 Dominion Circle, Los Angeles, CA 90040, (1983).

[2] Sibelius, Version 3.1, 20-22 City North, Fonthill Rd., London, N4 3HF

[3] Cubase, Steinberg, Steinberg Media Technologies GmbH., Neuer Hoeltigbaum 22-32, 22143 Hamburg, Germany

[4] Alty,J.L., Rigas, D., and Vickers, P., Using Music as a Communications Medium, *In Proc. Refereed Demonstrations, CHI'97 Conference on Human Factors in Computing Systems*, ACM Press, pp 89 – 96, (1997).

[5] Alty, J.L., Can we use Music in Human Computer interaction? *In People and Computers X*, Diaper, D., and Winder R., (eds.), Cambridge University Press, Cambridge, UK, (1995).

[6] Alty, J.L., and Rigas, D., Communicating Graphics by the Auditory Channel: An Empirical Approach, *Int. J. Human Computer Studies,* Vol. 62 pp. 21 – 40, (2005).

[7] Rigas, D. and Alty, J.L., The Rising Pitch Metaphor: An Empirical Study, *Int. J. Human Computer Studies*, Vol. 62 pp. 1 - 20, (2005).

[8] Rigas D., *Ph D. Thesis*, Loughborough University, UK, (1998)

[9] Vickers P., and Alty, J.L., *Interacting With Computers*, Using Music to Communicate Music Information, Vol. 14, No. 5, pp 435 – 446, (2002).

[10] Vickers P., and Alty, J.L., Using Music to Communicate Music Information, *Interacting With Computers*, Vol. 14, No. 5, pp 435 – 446, (2002).

[11] Vickers P., and Alty, J.L., Musical Program Auralisation: A Structured Approach to Motif Design, *Interacting With Computers*, Vol. 14, No. 5, pp 457 - 485, (2002).

[12] Vickers P., and Alty, J.L., When Bugs Sing, *Interacting With Computers*, Vol. 14, No. 6, pp 793 - 819, (2002).

[13] NASA Human Performance Research Group, “Task Load Index (NASA TLX) v1.0 computerised version”, *NASA Ames Research Centre*, (1987).

[14] Mayer-Kress, G., Bargar, R., & Choi, I. , Musical Structures in Data from Chaotic Attractors. In G. Kramer (Ed.), *Auditory Display* (Vol. XVIII, pp. 341-368). Reading, MA: Addison-Wesley. (1994)

[15] Quinn, M. (2000). Seismic sonata: a musical replay of the 1994 Northridge, California Earthquake.