

5. A structured approach to auralisation design

Men believe in the truth of all that is seen to be strongly believed in.

– Nietzsche

5.1. Introduction

In the previous chapter we discussed how a system was constructed to allow the general auralisation of Pascal programs at the construct level. After initial testing of the prototype, a set of auralisations based on semi-formal design principles was specified. To move towards a set of formal musical-auralisation design principles it is necessary to review issues surrounding the cognitive aspects of music and its use in the interface. In this chapter we discuss existing guidelines for the use of music in HCI and consider these in the light of work done in the music cognition and music-theoretic fields. We then propose a set of organising principles for musical auralisations. The chapter then contains a description of a new set of auralisations and discusses their evaluation through a study with advanced-novice programmers.

5.2. Existing guidelines

Some work has already been carried out to determine some general principles for using sound in the interface. Using the headings suggested by Brewster et al. [36], we will discuss some of these guidelines.

5.2.1. Timbre

Early studies on pitch perception and other psycho-acoustic phenomena tended to make use of sine waves. The recent advent of cheap, high-quality multi-timbral synthesisers has allowed researchers to use more complex sounds, or timbres. Alty [3] identified timbre as useful for distinguishing between different types of information, but people cannot easily distinguish the different timbres in chords. Brewster et al. [36] recommended using different harmonically rich timbres (as found in the sounds of musical instruments) for different classes of earcon to make them easy to distinguish. Common sense would suggest that certain timbres are easy to tell apart from each other (e.g. violin and piano). Rigas and Alty [137] carried out experiments to find which timbres and timbre classes work well as discriminating factors. Their study identified the following families of timbres:

Piano	Piano, harp, guitar, celesta, xylophone
Organ	Organ, harmonica
Wind	Trumpet, French horn, tuba, trombone, saxophone
Woodwind	Clarinet, English horn, pan-pipes, piccolo, oboe, bassoon, flute
Strings	Violin, cello, double bass
Drums	Drums

They stated: “*Our experiments suggest that one instrument from each of the following families is likely to be recalled by the listener with no prior training.*” [137]. The consequence of this is that there are really only six perceptibly unique timbres. One limitation of Rigas and Alty’s findings is that the timbres were generated by a low-quality synthesiser (Roland MT-32). It is quite possible that a synthesiser with more faithful reproductions of musical instruments would yield a larger set of useful timbres.

5.2.2. Register

Brewster et al. found that register, or octave positioning should not be used if absolute judgements were to be made between earcons [36]. If register has to be used then they recommended gaps of between two and three octaves to ensure a recognisable difference.

5.2.3. Pitch

Alty [3] suggested that there is a useful metaphoric mapping from a numeric domain to that of musical pitch. Pitch can be used to communicate shape and trends but cannot be used for accurate numerical determination. However, comparison of a pitch with a reference tone can be used for exact determination. Although the MIDI system can specify 128 unique pitches (not counting the microtonal intervals obtainable by pitch bending techniques), they are not all useable in HCI. A piano only has 88 pitches and the range that is useful in interface applications is smaller than that. According to Brewster et al. [36], pitches should lie in the frequency range of 125Hz to 5KHz (\approx B1 to \approx D#7).

5.2.4. Rhythm and duration

Adding rhythmic patterns to a sequence of notes increases the memorability of the sequence [3]. Brewster et al. [36] suggest making the rhythms of objects as different as possible. They also warn that very short notes may go unnoticed and so advise against using any notes shorter than 82.5ms (a semi-quaver triplet at a tempo of 120 beats-per-minute).

5.2.5. Intensity

Intensity, or perceived loudness is an important issue as (according to Brewster et al. [36]) "*it is the main cause of annoyance due to sound*". They recommend giving the user control over the volume and ensuring that all sonic objects lie within a narrow band of intensities so that none stands out from the others.

5.2.6. Spatial location

If possible, designers should make use of stereo positioning and three-dimensional sound placement (when available) [36]. Although location in the sound field can be very useful for helping to distinguish between audio streams, designers must keep in mind that not all users can benefit even from stereophonic placement. Users with hearing problems in one ear, or users without headphones are likely to lose any spatial cues. We would recommend that spatial information be used as additional clues rather than as a primary discriminator of sonic objects.

5.2.7. Musical scale

Another characteristic of music, and one that Brewster et al. did not address is that of musical mode or scale. We stated earlier that major and minor keys would be used to discriminate between Boolean *True* and *False*. Alty [3] asserts: “*Most people can recognise the differences between major, minor, pentatonic and whole-tone musical scales.*”

5.3. Review of cognitive aspects

There are certain cognitive aspects of music that may assist in the construction of music-based auditory displays. In a study of the perceptual organisation of tone sequences and melodies, Watkins and Dyson [155] stated:

“The perceptual organisation of tone sequences is facilitated if they are structured to approximate the schemata of melodies belonging to the musical idiom with which our listeners are familiar.”

In other words, if the music is in a style familiar to the listener then it is easier to recognise and understand. In the West, the schema with which we are most familiar is Western musical forms based upon the seven-note diatonic scale²¹. In fact, western musical forms are readily recognised around the world (especially in the computing community) [3] and so this gives a good starting point for auralisation design [153]. Watkins and Dyson go on to say:

“Tone sequences that obey the constraints imposed by the system of Western scales and keys are more easily organised [cognitively], learned and discriminated than control sequences of comparable complexity” [155].

5.3.1. Melody recall

One of the goals of this research was to show that people without formal musical training could make use of musical communication channels. Experimentation supports the idea that familiarity with tunes makes them readily discriminable from unknown music even by musically-unsophisticated listeners [155]. Therefore, there is no intrinsic requirement that users of an auralisation system possess formal musical training, merely that they become familiar with the tunes of the system.

²¹ The seven-note diatonic scale is perhaps best known by many through the Rodgers and Hammerstein song *Do-Re-Mi* sung by Julie Andrews in the musical film *The Sound of Music*.

In experiments on subjects' ability to recall melodies, Sloboda and Parker [145] observed that the most fundamental feature preserved in the recalled melody was its metrical structure. Further, musicians and non-musicians differed significantly only on one measure, that of the ability to retain the harmonic structure of the original. Therefore, it would be wise not to rely on ability to discriminate between harmonic structures in the auralisation motifs.

Lamont and Dibben [96] defined a motif as a “*core of pitch and rhythmic information which may be subjected to variation by a range of musical transformations*”. In music-theoretical terms, transformations are considered to be either:

- **Surface:**– where the changes are in terms of texture, orchestration, register, pace etc., or
- **Deep:**– e.g., derivation and fragmentation of the original pitch and rhythm information [96].

As variation of harmonic structure is a deep transformation, Sloboda and Parker are thus suggesting that the difference between musicians and non-musicians is that whilst musicians can recognise deep transformations, non-musicians are more likely to respond only to surface transformations. Lamont and Dibben [96] asserted that, according to the music-theoretical viewpoint, for tonal music, similarity relationships are in terms of musical motifs and surface transformations of the motifs create the differences.

Other studies cited by Lamont and Dibben have found that musically inexperienced subjects failed to identify the two themes of a piece of classical music having been misled by surface differences [61] and that after a single hearing of a classical tonal work, listeners' perceptions were influenced mainly by surface features such as loudness and texture [131].

However, Watkins and Dyson [155] observed that well-known melodies (e.g. nursery rhymes or folk tunes) are readily named by unsophisticated listeners. They cite this as an example of a musical skill that requires no formal musical training. But what it suggests is that even musically inexperienced listeners can distinguish between deep level transformations when the material is familiar. In a study on the perceived similarity of musical motifs, Lamont and Dibben [96] found that musically-trained listeners could make accurate similarity judgements between motifs on

the basis of deep transformations and that the surface variations did not confound their judgement.

Earlier we cited recommendations by auditory display researchers that timbre (a surface transformation) be used as a discriminant in sound design. We adopted this approach in the CAITLIN system and user responses indicate that it remains a useful tool (see appendix C). Such recommendations were made on the basis of short studies with no longitudinal aspects. From what we have learnt about music perception, it is possible that after prolonged exposure to such auditory interface components, these surface differences (timbre, register etc.) become invisible and the users begin to rely on the deep-level differences in the sonic elements. If surface characteristics are the only discriminating factors, then there is a possibility that auditory displays become confusing with time. Only longitudinal studies will answer this question.

5.3.2. Music contour

In the experiments described in chapter 4 subjects identified melodic contour as a useful aide-mémoire for recalling the motifs. Lamont and Dibben [96] also found that contour was one of the few surface characteristics that musically-trained listeners used when making similarity judgements between motifs. However, results reported by Edworthy [59] confirm Dowling's suggestion [55] that contour becomes important when tonal context is weak or confusing. Contour is less important in familiar melodies and melodies retained over a period of time. (The motifs in Lamont and Dibben's study were unfamiliar to the subjects.) The results from our preliminary experiments tend to bear this out. Whilst the motifs were designed according to specific principles, and thus motifs representing related constructs were based on the same melodic contour, this feature was of little concern to the listener. What is more important is that prior to using the CAITLIN system in earnest, subjects must first be exposed to the different motifs so that the tunes become familiar.

5.3.3. Transition probabilities

It has been observed that certain pitch combinations occur more often than others in western music [159]. Analysis of such pitch combinations gives us the notion of the transition probability [155] which describes the likelihood of a particular at-

tribute given an attribute in preceding elements. Listeners appear to be sensitive to transition probabilities that occur in western tonal melodies with commonly occurring transition probabilities being perceived as more musical” [155, 159].

Considering just the transition probabilities between the pitches of notes of equal duration, Watkins and Dyson [155] have shown that listeners are sensitive to those that reflect what is encountered in Western tonal music, judging melodies with commonly occurring transition probabilities as more musical [159].

Note	D \flat	A \flat	E \flat	B \flat	F	C	G	D	A	E	B	F \sharp
<i>q</i> value	-5	-4	-3	-2	-1	0	1	2	3	4	5	6

Table 5.1 *q* values of equal-tempered notes (from Watkins & Dyson [155])

The table shows, for the key of C major, the number of fifths (*q* value) between each note and the tonic of the key, i.e., C. For instance, we see that F is one fifth (seven semitones) below C, whilst D may be reached by ascending two fifths. In western music, for a given key, notes with absolute *q* values of six or more are unusual.

Each note in the equal-tempered scale is assigned a sharpness value (*q*) that represents the number of fifths (an interval of seven semitones) between it and its keynote (the root note of the scale, e.g. C in C major). For the key of C major, the *q* values of all twelve notes of the scale are shown in Table 5.1. The keynote acts as a tonal centre, or anchor point, in western tonal melodies. The difference between the *q*-values of two adjacent notes in a melody is called the “fifth-span” (or δq) which is given in **(1)** as the absolute value of the difference in *q*-values of the two notes of the interval [101, 102].

$$\mathbf{dq} = Abs(q_1 - q_2) \quad \mathbf{(1)}$$

Watkins and Dyson [155] discovered from an analysis of a large number of existing melodies:

“... that successive fifths spans tended to be less than 6 fifths. Fifth spans larger than this were rare. Any large fifth spans that did occur were not found in adjacent intervals. These observations suggested that such constraints may provide listeners with information about the key of a melody: Each interval will suggest a range of likely keynotes, for which the fifth span of the interval is less than 6. Each interval will also suggest a range of unlikely keynotes, for which the fifth span is greater than 6. In this way, if the listener looks at the melody in terms of this progression of fifths, then the keynote will appear as an orienting influence amidst the dancing note patterns.”

Looking at Table 5.1 it can be observed that the diatonic scale consists of intervals with a fifth span of five or less. Thus, for the typical western listener, intervals will be expected to lie within the diatonic scale.

Studies on similarity judgements and internal representation of musical structure [51, 96] have shown that tonal music is more easily organised cognitively and more easily recognised than atonal music. Restricting musical auditory displays to tonal schemes would, therefore, be advantageous.

5.3.4. Hierarchic structural cognition

Dibben [51] found evidence that people internally represent tonal musical in terms of a hierarchy of events. The theory of hierarchical organisation was put forward by Heinrich Schenker [143]. Schenker discovered that the surface events in tonal music are specifically related to a fundamental hierarchical organisation [51, 143], that of the tonic triad. Deutsch and Feroe [50], Lerdahl and Jackendoff [99], Narmour [124] and Swain [149] proposed that hierarchical structure is the way that listeners hear and represent the structure of tonal music [51]. Using a two-dimensional visual mapping of the octave, Holland's [78] Harmony Space system enables users to learn complicated concepts about music in a short space of time. By mapping four notes of the octave to the horizontal axis of a space and the other three notes to the vertical axis, many musical structures show up as regular patterns.

5.4. Organising principles

Using the knowledge gained from the heuristic approach described in chapter 4 and the findings of the research described above it is possible to set out some general guidelines to be followed when constructing musical auralisation motifs. We suggested some preliminary guidelines elsewhere [153]. The following guidelines are a development and refinement of those earlier suggestions.

1. **Tonality:**— Many early auditory displays made arbitrary mappings between data and pitch (or even worse, between data and frequency). Music is much more easily organised when it uses a familiar schema. Because of the ubiquitous nature of Western music (particularly pop music), the diatonic scale and its close relations (minor and pentatonic) should be used wherever possible. Tran-

sition probabilities show that intervals outside the diatonic scale are less likely to occur in Western music and so such intervals should be reserved for specific purposes (such as when attention needs to be drawn to an event). Specific modalities can be useful for particular mappings. For instance, the major and minor modes are convenient metaphors for the Boolean values *True* and *False*.

2. **Hierarchy:**– Organise the motifs in a hierarchy based on the taxonomy of target language constructs (e.g. Figure 4.10). If representing data structures musically, then mappings should be chosen that make use of our tendency to cognitively organise the music hierarchically.
3. **Structure:**– Of the deep characteristics, metre and rhythm are more easily retained than harmonic structure [36, 145]. Therefore, the rhythm of a motif will play a more important role in listener discrimination than its melodic and harmonic features. However, melody still plays an important role; with repeated exposure melodies become internalised and easily recognisable regardless of surface transformations such as altered timbre.
4. **Surface features:**– Care should be taken when relying on the surface features of music to convey information. To untrained listeners surface features such as timbre, register, pace and contour are most used to form discrimination judgements. Indeed, researchers have identified timbre as a particularly useful discriminant [3, 28, 36, 135, 137]. However, as music becomes more familiar (as would be the case with earcons and auralisation motifs after repeated use), listeners tend to make more use of deeper features such as metre and rhythm. For instance, nursery rhymes are easily recognised regardless of speed, register, and timbre. These surface features should be used to provide clues that make initial learning of the system easy. However, for long term use, motifs should also be markedly different rhythmically and harmonically to ensure that when changing from being novice to expert system users, the music is still distinguishable once reliance on surface clues diminishes.
5. **Percussion:**– Composers use percussion alongside the tuned instruments to enhance the music and to add emphasis at particular points in the score. Percussive devices should be used as extra cues in motifs to help users recognise significant events (such as a change in condition of a Boolean expression).
6. **Prolongation:**– In the program domain (less so in user-interface applications), information about continuous state is needed for proper comprehension.

In a program the entry and exit points of the constructs are separated by the various Boolean expressions and statement blocks contained within. To prevent loss of context in the listener's mental model of the program *drones* should be incorporated into all construct motifs. The drones should start playing upon entry to the construct and should continue playing until the program exits from construct. Of necessity timbres that permit indeterminate duration (such as string, organ and wind instruments) should be used for the drones. To facilitate recognition of nested constructs, the pitch of a construct's drone should be related to its nesting depth, and intervals between drones should be in keeping with the overall key and tonality of the auralisation.

To summarise, auralisations should be based around diatonic tonalities and should be organised hierarchically. They should make good use of surface level discriminants (e.g. timbre, register, pace, and contour) and deep-level discriminants (e.g. metre and rhythm) to allow identification by new users and by experience users alike. Percussive devices should be used alongside the pitched messages to provide emphasis at key points. Finally, auralisations should provide continuous information about persistent features (e.g. by the use of drones).

5.5. Redesigned motifs

Taking the above into consideration, a final set of auralisations was constructed. However, to explore the effect that harmonic structure has on the perception of difference, the level 3 tunes were differentiated more by harmonic features than by rhythmic features. For example, looking at Figure 5.3 we observe that the rhythmic structure of the **WHILE** and **REPEAT** loops are the same. The two loops differ only in their melodic and harmonic structure (the **WHILE** has a harmonic progression of I-V-Vm-Im and the **REPEAT** has a progression of Vm-IVm-IV-I²²) and their timbres. Experimentation should show whether the deep or surface features had the most impact.

The final set of auralisations is shown in Figure 5.2 (CD tracks 21-28) and Figure 5.3 (CD tracks 29-32). Notice that all constructs are shown with the appropriate

²² In music theory, degrees of a scale are written as I for the tonic, or root, up to VII for the seventh note. If a chord is minor then the roman numeral is suffixed with an 'm'.

auditory parentheses and that all the selections now have drones (prolongation). Percussion is used to signal key points in program flow. For instance, the auditory parentheses are percussive as are the terminal conditions in the iterations. All motifs are in a diatonic scale (with accidentals given for emphasis). The hierarchy of Pascal constructs has been preserved in the motif design. Differences between the motifs are provided by timbre and contour (surface characteristics) and, to a lesser extent, by rhythm (deep-level). Using these new motifs, the following Pascal program,

```
PROGRAM Exemplar ;
VAR
  counter : Integer ;
BEGIN
  counter := 1 ;
  WHILE counter <= 2 DO
  BEGIN
    IF counter MOD 2 = 0 THEN
    BEGIN
      Writeln ('Counter is even') ;
    END ;
    counter := counter + 1 ;
  END ;
END.
```

would generate the auralisation shown in Figure 5.1 (also on CD track 20).

Auralisation score of a program. Notice that the drones for each construct appear together on the staff labelled 'Drones'. Likewise, all percussive events for each construct appear on the 'Drums' staff. Shifts between major and minor are shown by changes in key signature.

©1999 Paul Vickers
All rights reserved

WHILE

IFs

Drones

Drums

iteration open

WHILE drone

selection open

True

IF drone

False

counter := 1

WHILE condition

selection close

IF drone

True

counter := counter + 1;

False

exit WHILE

exit IF

closed triangle

iteration close

Writeln... counter := counter + 1;

sleighbell denotes terminating condition for WHILE

Figure 5.1 Score for a complete program (Trk 20)

IF yielding True

IF yielding False

IF...ELSE yielding True

IF...ELSE yielding False

CASE match

CASE no match

CASE...ELSE match

CASE..ELSE no match

auditory parentheses for selection: - rising and falling cowbell sounds

Figure 5.2 Final auralisations – selections (Trk 21-28)

WHILE

auditory parentheses for iteration: - open and close triangle timbres

REPEAT

FOR...TO

6 iterations

drone

sleighbell on last iteration

FOR...DOWNTWO

Figure 5.3 Final auralisations – iterations (Trk 29-32)

5.6. Auralisation recognition study

5.6.1. Objective

The objective of this first study was to investigate how novice programmers would interpret the redesigned auralisations. Twenty-two subjects took part in the investigation. Additionally, as the system is intended to be usable by programmers regardless of musical expertise, it is hoped that the study would show no significant difference in performance across subjects with varying levels of musical knowledge and experience.

5.6.2. Subjects

Twenty-two subjects took part in the study. All were undergraduates on computing courses at Loughborough University. The experiment was carried out as part

of a second-level undergraduate course in human-computer interaction. The subjects were each paid £10.00 for their participation.

21 of the subjects were male, although this gender imbalance is typical of undergraduate computing courses in the UK. The mean age of the subjects was 21 (min. 19, max 23). None of the subjects reported any problems with their hearing.

5.6.2.1. Musical background

Musical knowledge and experience was measured by four variables: *interest*, *play*, *sing* and *musical score*. The *interest* variable (question 7 on the questionnaire in Appendix B) attempts to measure the general interest of the subject in music. The responses were scored using the following scale:

0. No interest in music at all
1. Enjoy listening to music
2. Enjoy performing music (alone, with friends or professionally)
3. Enjoy listening and performing

The *play* (question 8) variable is a simple Boolean flag stating whether or not a subject plays a musical instrument.

Sing (question 9) is a measure of how much subjects participate in singing. The possible values are:

0. I do not sing
1. I sing in the bath
2. I sing informally to others
3. I sing in a choir
4. I sing semi-professionally
5. I sing professionally

Musical score represents the score attained by subjects on the musical knowledge test of the questionnaire (questions 10 through 14). One mark was given for each correct answer resulting in a range 0 to 15.

5.6.2.2. Descriptive statistics

We observe from Chart 5.1 that no students reported having no interest at all in music, with the majority (14) stating that they enjoy listening to music. The remaining eight subjects enjoyed performing music. It is strange that of these eight, five

claimed to enjoy performing but not listening to music. It is possible that these subjects misread the instructions on the questionnaire and thought they must only tick one box when, in fact, they were invited to tick all boxes that applied.

The *play* variable showed that eight subjects played an instrument, a result that is consistent with the responses to the *interest* question.

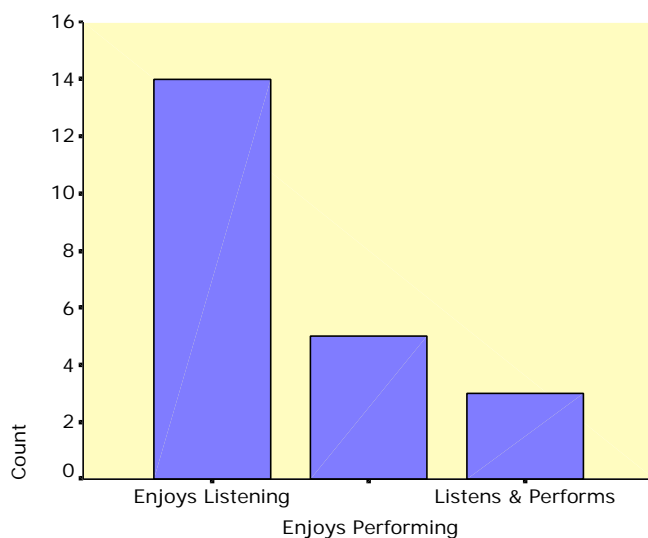


Chart 5.1 Distribution of *interest* variable

This chart shows that 14 subjects enjoyed listening to music, 5 enjoyed performing music, and 3 both listened to and performed music.

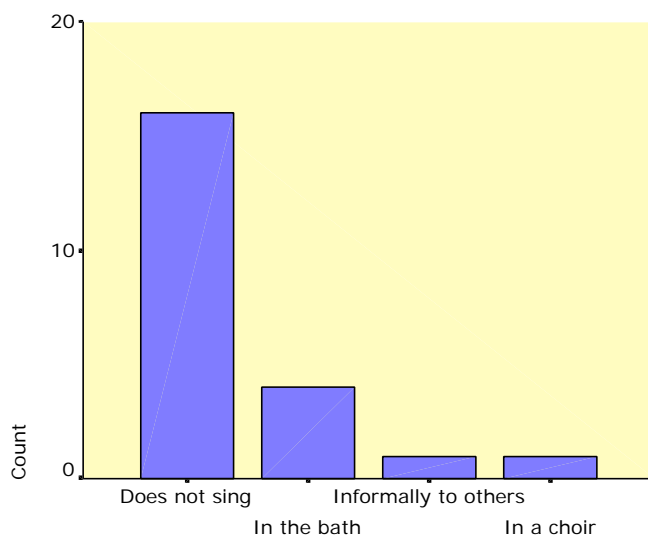


Chart 5.2 Distribution of *sing* variable

The majority of subjects reported that they did not sing. A minority sang at various informal and amateur levels.

Most of the subjects (16) did not claim to be singers (see Chart 5.2). None of the remainder sang above the amateur level.

One would not expect computer science undergraduates to have extensive knowledge of musical terms and theory and the results of the musical knowledge test bear this out. From the bar chart below (Chart 5.3) we see the majority (19) scored 6 or less on the knowledge test. The three subjects that scored 12 or above played musical instruments as well.

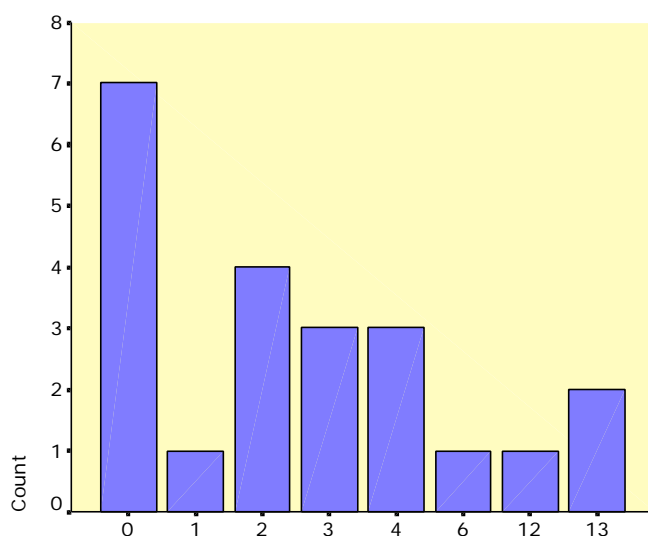


Chart 5.3 Distribution of *musical knowledge* scores

The abscissa of this chart shows the scores attained by subjects on the musical knowledge test. The maximum possible score was 15. The ordinate shows how many subjects achieved the various scores. Hence we see that seven subjects scored zero, two scored 13 and none scored 5.

Eight subjects played musical instruments. Six subjects claimed to enjoy singing. None of the subjects reported having no interest at all in listening to music. All subjects reported a western-style cultural upbringing.

Because we were interested in how musical knowledge and experience might affect the ability to make use of the auralisations, subjects were given a short test of musical terms and concepts. The maximum possible score was 15. Chart 5.4 shows the distribution of scores of the twenty-two subjects. The three subjects with the high scores all played musical instruments too.

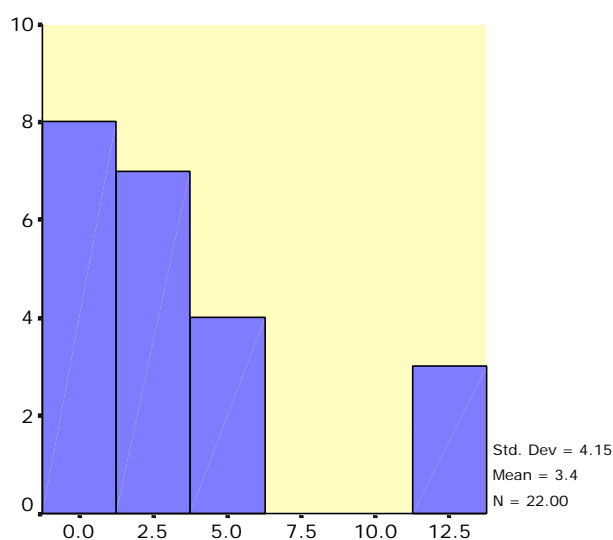


Chart 5.4 Musical Knowledge Test

This chart shows the same musical knowledge scores, but this time as a statistical distribution. This representation shows the large gap in scores between the majority of subjects and the cluster of three who scored between 12 and 13.

The box plot in Chart 5.5 shows how musical knowledge scores were distributed amongst those who played an instrument and those who did not. Not surprisingly, the scores amongst the instrument players were higher. An independent samples Mann-Whitney test (see Table 5.2) showed the difference in scores between the two groups was significant ($p=0.019$).

Test Statistics^b

	Musical Knowledge
Mann-Whitney U	22.500
Wilcoxon W	127.500
Z	-2.336
Asymp. Sig. (2-tailed)	.019
Exact Sig. [2*(1-tailed Sig.)]	.020 ^a

a. Not corrected for ties.

b. Grouping Variable: Plays Instrument

Table 5.2 Comparison of musicians and non-musicians

The non-parametric Mann-Whitney test shows that players of musical instruments scored significantly higher on the musical knowledge test than non-instrument players.

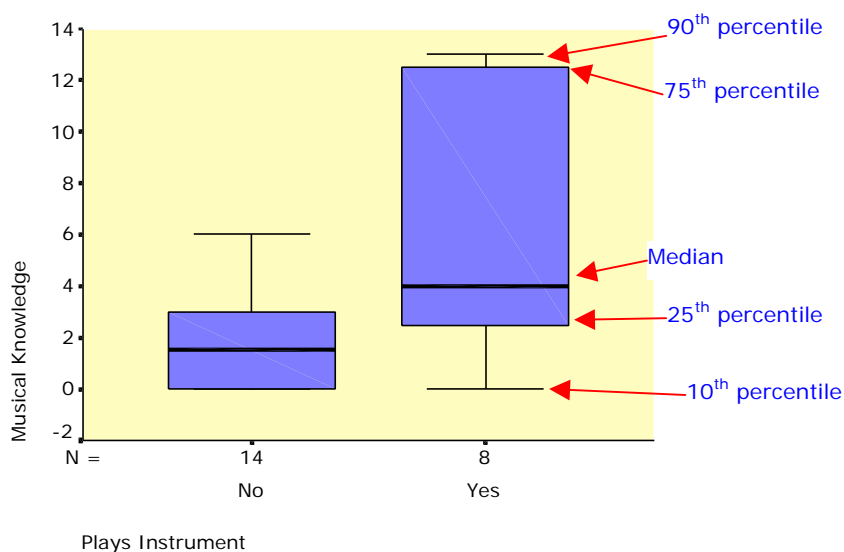


Chart 5.5 Musical scores of musicians vs. non-musicians

The box plot shows that the 14 subjects who did not play musical instruments scored low on the musical knowledge test and that their scores were quite bunched. By contrast the 8 musicians scored much higher and had a greater spread in their scores.

5.6.2.3. Programming experience

All subjects had an average of two years' programming tuition and had all written programs in Pascal. On average, the subjects had written programs in five programming languages (min. 2, max 7).

5.6.3. Task

The session was divided into two tests. Test 1 required subjects to listen to and identify forty construct auralisations. Test 2 involved listening to pairs of auralisations which were either nested, or sequential. In all, sixty construct auralisations were used in the two tests. Half of the auralisations were of iteration constructs, half of selections. Auralisations were of differing lengths, depending on the construct type. The auralisations were generated by the CAITLIN system using a Boss DS-330 multi-timbral synthesiser and were played to subjects over a stereo amplifier and speakers in a tiered lecture theatre. The volume was adjusted so that the music could be heard well at the back of the room. The speakers were placed relatively close together so as to mask any stereo effects.

Subjects were provided with worksheets on which they recorded their responses. For Test 1 there were forty worksheets; for Test 2 there were ten worksheets.

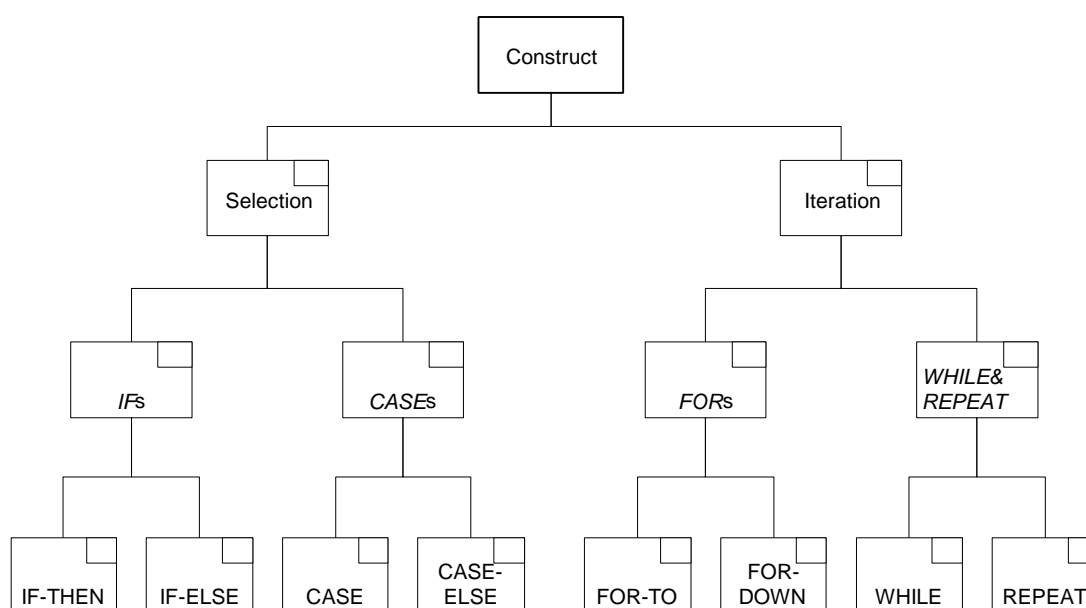


Figure 5.4 Construct identification worksheet

Subjects were asked to identify a construct auralisation by ticking one box on the chart. If they were unable to identify a construct exactly, they could tick a higher-level box on the chart.

In both tests, subjects recorded their response by ticking a box on a chart, each box representing a specific construct (see Figure 5.4). In test 2, there were two box charts per exercise and an additional box in which subjects could state whether they heard sequential or nested constructs.

5.6.4. Procedure

The study was subdivided into four sessions:

1. Introduction and tutorial.
2. Auralisation drilling & feedback.
3. Listening test 1.
4. Listening test 2.

All four sessions ran contiguously (apart from a short break between sessions 3 and 4). At the beginning of the experiment each subject was given a workbook²³ which contained:

- The full written text of the introduction and tutorial session.
- A subject experience questionnaire.

²³ Appendix B contains a copy of this workbook (with all but one each of the worksheets removed).

- The worksheets for the individual exercises.
- A space for detailed subject responses and general feedback on their experience of the experiment (the responses are reproduced verbatim in Appendix C).

5.6.4.1. Introduction and tutorial

The subjects had no prior experience of program auralisation and so this briefing session was given. The tutorial involved explaining the philosophy of the technique and giving examples of the various auralisations used. An explanation of how the auralisations are constructed and how they represent the various components of the constructs was given.

5.6.4.2. Auralisation drilling and feedback

Next, a practice test was run. Twenty auralisations were presented to the subjects. Each was played three times, during which time subjects wrote down what construct they thought was being represented. After each exercise, immediate feedback was given so that subjects could check their answers. Following this the subjects completed the subject data questionnaire which was used to gather information about their musical background and programming experience.

5.6.4.3. Listening test 1

The third session was the first of two actual listening tests. Prior to commencing the test subjects were given instructions on how to fill in their worksheets. Opportunity was given for questions to be asked.

The listening test itself involved the presentation of forty auralisations. For each, subjects had to tick the box on the worksheet that they thought matched the auralisation being played. Each auralisation was played three times with a short pause between repetitions.

The forty auralisations were made up of ten unique selection and ten unique iteration auralisations. These were then randomly arranged into two sets of twenty, each unique auralisation occurring once per set.

Because of the hierarchical nature of the constructs and the corresponding auralisations, the worksheets allowed subjects to give different levels of answer. For instance, if a subject thought an auralisation represented a *FOR...TO* loop then they would tick the box labelled *FOR-TO*. This is the highest level of identification which

we have previously called *specific identity*, that is, identification of the precise construct is achieved. If the subject could identify the loop as a `FOR` but was unsure whether it was a `FOR...TO` or a `FOR...DOWNTO` then they could tick the higher level box labelled *FORs*. This is identifying the construct at its *sub-class* level. The lowest level of identification, the *class* level is where, in this instance, the subject can identify the auralisation as a loop as opposed to a selection, but is unable to be any more specific. This would involve the subject ticking the *Iteration* box on the worksheet. Upon completion of the listening test a short break was given.

5.6.4.4. Listening test 2

The second listening test was similar to the first except that this time twenty auralisations were presented in ten pairs. The reason for playing pairs of auralisations was to see whether subjects could correctly discriminate between sequential and nested constructs. That is, would the auralisations allow subjects hear the difference between say, an `IF` nested within the body of a `WHILE` loop and a `WHILE` loop followed by an `IF`?

Twenty unique auralisations were used, ten iterations and ten selections. Five pairs were nested, the other five were sequential.

As in the first test subjects were asked to tick boxes on worksheets to identify the constructs they heard. This time each worksheet had two response charts labelled *First Construct* and *Second Construct*. These charts were used to identify the first and second constructs of each pair respectively. Additionally, two further boxes were provided, labelled *Nested* and *Sequential*, which were used to identify whether the constructs were nested or sequential (see Appendix B). Each pair of constructs was played three times.

5.6.4.5. Subject feedback

Following the two listening tests, subjects were invited to write down their responses to the experiment on a page in the workbook. These responses are reproduced in full in appendix C.

Finally, subjects handed in their workbooks, after which they received their payment.

5.6.5. Results

The subjects' responses were scored and analysed to see how they interpreted the auralisations. Earlier we stated that constructs could be identified at three levels:

- *Specific identity*: the construct is identified exactly.
- *Sub-class*: the construct is identified at the level of its siblings (e.g. *IFs* is the sub-class of both *IF* and *IF...ELSE*).
- *Class*: the construct is identified merely as an iteration or a selection.

By assigning subject responses to one of these categories, it is possible to discover at what level of identity subjects could recognise constructs the best.

5.6.5.1. Listening test 1

The first listening test presented forty auralisations to twenty-two subjects, making a maximum of 880²⁴ discrete observations. The results for the test are summarised in Table 5.3 which shows the combined scores of subjects at the various levels of construct identity for each of the forty exercises. For example, reading from Table 5.3 we notice that for the first auralisation, 23% of subjects correctly identified the construct as a **REPEAT** statement; a further 27% who did not achieve the highest level of accuracy were able to identify it as an unbounded loop (**WHILE** or **REPEAT**) and 41% at least identified the construct as an iteration. The remaining 9% misidentified the construct as some form of selection.

The mean identification of *specific identity* by subjects was 46%. Identification at the *sub-class* level ran at 30%. The mean score for identification at *class* level was 21%. This leaves a mean absolute misidentification rate of 3%.

5% of the iterations (22 out of 440 iteration observations) were misidentified as selections. Subject 19 accounted for 10 of these errant responses. None of the selections was incorrectly identified as an iteration (the reason that the responses to exercise 28 do not total column of 100% is because subject 16 did not give an answer).

²⁴ The actual number of observations is 877 as three subjects each left an exercise sheet blank.

No.	Construct	% Correct				No.	Construct	% Correct			
		SI	SC	C	Tot.			SI	SC	C	Tot.
1	REPEAT	23%	27%	41%	91%	21	CASE...ELSE	14%	82%	5%	100%
2	IF	41%	32%	27%	100%	22	WHILE	45%	41%	9%	95%
3	FOR...TO	45%	9%	32%	86%	23	IF	50%	36%	14%	100%
4	CASE...ELSE	59%	32%	9%	100%	24	FOR...TO	73%	9%	18%	100%
5	FOR...DOWNT0	59%	9%	27%	95%	25	CASE	50%	41%	9%	100%
6	IF...ELSE	18%	14%	68%	100%	26	REPEAT	50%	32%	14%	95%
7	CASE...ELSE	5%	32%	64%	100%	27	CASE...ELSE	86%	14%	0%	100%
8	WHILE	27%	41%	23%	91%	28	IF	32%	55%	9%	95%
9	CASE	36%	5%	59%	100%	29	FOR...TO	59%	5%	23%	86%
10	IF	55%	5%	41%	100%	30	WHILE	45%	41%	5%	91%
11	FOR...TO	55%	0%	41%	95%	31	CASE...ELSE	9%	36%	55%	100%
12	CASE	50%	41%	9%	100%	32	IF...ELSE	14%	27%	59%	100%
13	WHILE	27%	50%	9%	86%	33	FOR...DOWNT0	64%	14%	18%	95%
14	FOR...TO	100%	0%	0%	100%	34	WHILE	55%	45%	0%	100%
15	REPEAT	73%	27%	0%	100%	35	CASE	50%	14%	36%	100%
16	CASE	59%	23%	18%	100%	36	FOR...TO	100%	0%	0%	100%
17	IF...ELSE	32%	45%	23%	100%	37	REPEAT	41%	50%	9%	100%
18	WHILE	23%	73%	5%	100%	38	CASE	59%	27%	14%	100%
19	CASE...ELSE	5%	91%	5%	100%	39	FOR...DOWNT0	55%	18%	14%	86%
20	FOR...DOWNT0	77%	5%	14%	95%	40	IF...ELSE	32%	41%	27%	100%
Mean 1-20		43%	28%	26%	97%	Mean 21-40		49%	31%	17%	97%
Mean 1-40		46%	30%	21%	97%	Key: SI=Specific Identity, SC=Subclass, C=Class					
						!Subject 13 did not answer this exercise					
						*Subject 16 did not answer this exercise					
						#Subject 14 did not answer this exercise					

Table 5.3 Identification of constructs, test 1

For each of the forty constructs the mean correct scores at the three levels of identification are given.

5.6.5.2. Listening test 2

In the second listening test ten pairs of auralisations were played to twenty-one subjects (subject 15 did not participate). Five of the pairs were of nested constructs, the other five of sequential constructs. This gives 420 auralisation observations, 105 nesting results and 105 sequential observations (there were no missing data). A summary of the results is given as Table 5.4. The results are comparable with those for the first listening test, there being no significant difference between the two in terms of auralisation identification rates. It is interesting that the same results were obtained even though in this test subjects were required to identify two constructs-per-exercise rather than one. Furthermore, in half the cases the constructs

were nested which meant that subjects were listening to interleaved auralisations. That this did not impair performance is encouraging.

No.	Construct	Type	SI	SC	C	N/S
1a	REPEAT	Nest	48%	19%	29%	95%
1b	IF...ELSE		38%	43%	19%	
2a	WHILE	Nest	38%	38%	10%	95%
2b	IF...ELSE		29%	57%	14%	
3a	REPEAT	Seq.	33%	43%	5%	100%
3b	IF		38%	38%	24%	
4a	REPEAT	Seq.	33%	48%	14%	100%
4b	CASE		71%	5%	24%	
5a	CASE...ELSE	Nest	14%	57%	29%	100%
5b	FOR...TO		52%	19%	19%	
6a	FOR...DOWNT0	Seq.	95%	5%	0%	100%
6b	IF...ELSE		24%	48%	29%	
7a	FOR...DOWNT0	Nest	90%	10%	0%	95%
7b	FOR...TO		90%	10%	0%	
8a	FOR...TO	Nest	90%	10%	0%	100%
8b	IF		24%	67%	10%	
9a	IF...ELSE	Seq.	29%	57%	14%	95%
9b	WHILE		33%	57%	10%	
10a	CASE...ELSE	Seq.	33%	52%	14%	100%
10b	CASE		67%	14%	19%	
Mean			49%	35%	14%	98%

Key: SI=Specific Identity, SC=Subclass,
C=Class, N/S=Nest/Sequence

Table 5.4 Identification of constructs, test 2

The table shows the mean correct scores at each level of identification. In addition, each construct pair is labelled as either nested or sequential.

The additional task in this test was to identify whether the two constructs heard in the auralisation were sequential or nested. The results of this aspect are particularly encouraging with Table 5.4 showing a mean identification rate of 97.5% correct. Of the 105 subject-nest responses, only three were incorrectly identified as sequences; 1 of the 105 subject-sequence responses was misidentified as a nesting.

5.6.6. Discussion

5.6.6.1. Musical knowledge and experience

One of the motivations behind this research was to develop a musical auralisation system that does not require any special musical skills, knowledge or experi-

ence. From the data collected in the questionnaire we can see whether musical experience had any effect on subjects' performance.

5.6.6.2. Effect of musical knowledge and experience.

The CAITLIN system was intended to be useful to programmers regardless of their musical expertise. We notice that the four musical variables are not orthogonal factors because there are interactions between them. For instance, the score given to *interest* is influenced by whether the subject plays a musical instrument (see Chart 5.1 above). We also notice that there is a significant correlation between subjects' scores on the musical knowledge test and whether or not they play an instrument (see Table 5.2 and Chart 5.5 above). For this reason there is little to be gained from running individual tests for each variable. Therefore, the data were analysed by multiple-linear-regression to see whether any or all of the musical experience factors had any influence on the scores. For each of the two listening tests the four musical variables were analysed for their effect on the subjects' correct scores at the *specific identity* and *sub-class* levels. The results of the tests are given below in Table 5.5 through Table 5.8 and show that musical knowledge and experience had no significant effect on subjects' ability to recognise constructs at the *specific identity* or *sub-class* levels. In fact, the best predictor of a subject's score is to take a value approximating the mean of the sample.

Coefficients ^a

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	17.414	8.539		2.039	.057
Musical Knowledge	.225	.668	.104	.337	.740
Plays Instrument	-2.531	12.227	-.139	-.207	.838
Musical Interest	1.063	7.807	.088	.136	.893
Sings	-.840	3.518	-.075	-.239	.814

a. Dependent Variable: Test 1: Specific Identity

Table 5.5 Regression test for effect on *Specific Identity* scores (test 1)

The multiple linear regression shows the probability of each of the listed factors having an effect on the subjects' performance. From this table we see that none of the musical factors had a significant effect on subjects' identification of constructs. In fact, the best predictor of the score was the regression model's constant term that had a probability of 0.057.

From Table 5.5 we observe that the constant term (17.414) in the regression equation ($y = 17.414 + (.225 \times \text{musical score}) - (2.531 \times \text{play}) + (1.063 \times \text{interest}) - (.840 \times \text{sing})$) is the best predictor of a subject's score ($p=0.057$). None of the four factors are significant predictors of correct score ($p=0.740... 0.893$).

The result is mirrored in the rest of the tests except in Table 5.7 where not even the model constant could be used as a predictor of subjects' scores.

Coefficients^a

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	16.408	5.551		2.956	.009
Musical Knowledge	-.423	.434	-.291	-.974	.344
Plays Instrument	.654	7.948	.053	.082	.935
Musical Interest	-.397	5.075	-.049	-.078	.939
Sings	1.577	2.287	.208	.690	.500

a. Dependent Variable: Test 1: Sub Class

Table 5.6 Regression test for effect on *Sub-class* scores (test 1)

A further regression model tests the effects of the musical variables on the sub-class identification scores. Again, only the model constant had any predictive ability.

Coefficients^a

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	4.967	4.447		1.117	.280
Musical Knowledge	.225	.348	.190	.647	.527
Plays Instrument	-2.228	6.368	-.223	-.350	.731
Musical Interest	3.352	4.066	.505	.824	.421
Sings	-1.633	1.832	-.265	-.891	.385

a. Dependent Variable: Test 2: Specific Identity

Table 5.7 Regression test for effect on *Specific Identity* scores (test 2)

The regression model for musical factor effect on specific identity scores for test 2 again reveals no effect.

Coefficients ^a

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	9.286	3.190		2.911	.010
Musical Knowledge	-.197	.249	-.240	-.789	.441
Plays Instrument	2.717	4.568	.393	.595	.560
Musical Interest	-1.483	2.917	-.322	-.508	.618
Sings	.272	1.314	.064	.207	.838

a. Dependent Variable: Test 2: Sub Class

Table 5.8 Regression test for effect on *Sub-class* scores (test 2)

The sub-class scores on test 2 were not affected by the musical factors according to the regression model.

5.6.6.3. Specific identity scores

From the preliminary study described in chapter 4 through the pilot studies to this experiment we observe a decline in subjects' ability to use auralisations to identify constructs at the level of specific identity (see Table 2.1). There are several possible reasons for this, the most obvious being that the earlier studies had smaller sample sizes (typically eight subjects) in which the subjects also had more programming experience.

Preliminary	78%	14%	1%	93%
Pilot 1	54%	25%	7%	86%
Pilot 2	57%	15%	9%	81%
Full study	46%	30%	21%	97%

Table 5.9 Comparison of test means

The mean correct scores at the three levels of identification are compared for the four experiments: the original preliminary study, the two pilot studies and the full study described above.

If subjects were simply guessing the specific identity of the constructs, then we could expect one out of every eight responses (12.5%) to be correct, there being eight constructs from which to choose. If we assumed that subjects were guessing then we would expect half the specific identity scores to be greater than 12.5% and half to be less. Looking at Table 5.3 we observe that for listening test 1, three specific identity scores (no.'s 7, 19, and 31) were less than 12.5% and the remaining 37 were higher. This is a highly significant difference ($\chi^2=15.62$, 1df, $p<0.01$). Considering the lack of confusion between iteration and selection classes, a more realistic approach would

be to say that for any given auralisation, we may assume that subjects know whether it is an iteration or selection. This narrows the choice to one amongst four. In this case we observe that seven exercises had specific identity scores less than the 25% attainable by guessing. Again, we see that these specific identity scores are still significantly higher than might be expected by guesswork alone ($\chi^2=8.05$, 1df, $p<0.01$).

Analysis of the selections and iterations separately reveals where the main difficulties with identification lay. If we again assume that subjects could distinguish between iterations and selections (recall that identification at the class level was 97%) then for any of the twenty selection auralisations, subjects could get one in four of them correct by guesswork. Again, if this were the case we would expect half (ten) of the specific identity scores to be greater and half to be less than 25%. Table 5.3 shows that six of the twenty selections had scores of 25% or less at the specific identity level and fourteen had scores higher. This difference is not significant ($\chi^2=0.94$, 1df, $p>0.1$). For the iterations, there is no evidence for guesswork ($\chi^2=5.83$, 1df, $p<0.02$). The six selection constructs with the low scores were the IF...ELSE and CASE...ELSE. This suggests that the motifs for the selection constructs' opening point-of-interest were not sufficiently different from each other. We must be cautious, however, not to conclude that guesswork alone was involved in selection identification. A low score does not necessarily mean that subjects incorrectly identified a construct. For instance, in exercise 6, five subjects only attempted identification at the sub-class or class levels.

It is interesting that less confusion surrounded the REPEAT and WHILE loops given that their surface characteristics were very similar. It is possible that the extra clue given by the placement of their conditional evaluations was sufficient to aid identification. For instance, the WHILE loop begins in a major mode and turns minor when the condition goes *False*; the REPEAT loop works in the opposite manner. Subject 2 commented: "*Listened for major and minor tones to distinguish between REPEAT and WHILE*" (see appendix C).

Furthermore, subjects were not faced with just eight choices, but had the opportunity to select higher nodes in the hierarchical charts. Analysis of the sub-class scores further suggests that more than guesswork was involved. A construct could be identified at its sub-class level by one of two ways:

- Either, the subject *deliberately* selects the sub-class box on the score sheet,
- Or, the subject misidentifies the construct as its sibling, resulting in a sub-class identification by default.

In the first listening test, of those observations scored correct at the sub-class level (261 in all), 59% were deliberate, 41% by default. This means that of those constructs identified at the sub-class level, just under half were as a result of confusing a construct with its sibling. We may conclude that subjects were generally understanding what they were hearing, but that at the specific identity level construct auralisations were still not distinct enough from their siblings.

Although there is no direct evidence to support it, another theory might be proposed to do with Miller's seminal work on information processing [118]. Miller identified that the channel capacity, or the greatest amount of information that can be given by an observer about various stimuli on the basis of an absolute judgement, is 2.5 bits²⁵; that is, people are able to discriminate between six ($2^{2.5} \approx 6$) equally likely alternatives when presented with values of one-dimensional data (e.g. musical pitches). As the dimensionality of the stimuli increases, then so does the channel capacity.

This experiment required subjects to make absolute judgements about eight stimuli. An absolute judgement is where the stimulus is presented and the observer is asked to identify it immediately. Given that relatively little training and learning of the system was provided, subjects had not had time to develop a good semantic-level understanding of the motifs (see section 5.6.6.4 below). Therefore, absolute judgements were being made between four sub-classes of motif (at which level we observed a mean correct response of 76%) and between eight individual constructs arranged as four pairs of motifs whose differences were slight compared with the sub-class differences.

Compare the results of this study with those of the preliminary experiment described in section 4.3 which achieved a mean sub-class identification score of 92% and a mean individual construct identification score of 78%. Examination of the motifs used in the preliminary study offers a possible explanation of these differences.

²⁵ In this context the *bit* is taken to mean the amount of information needed to make a decision between two equally likely alternatives.

In the preliminary study the motifs were arbitrarily-assigned tunes:

- REPEAT – twinkle twinkle little star
- WHILE – a plagal cadence (the amen at the end of a hymn)
- FOR – three notes on a piano that say ‘dah dah dit’ followed by an upwards or downwards scale depending on the direction of the loop followed by the closing ‘dit dit dah’.
- IF & IF...ELSE – the ‘fog horn’
- CASE – not used in this study

In effect, there were only four different motifs to memorise, and those motifs were quite different. No attempt was made to model the sub-class level, although, serendipitously, the four sub-classes were quite distinct. Therefore, the level of absolute judgement required of the subjects was much lower (2 bits of information).

5.6.6.4. The role of context

It is important to note that in this experiment the information presented to the subjects was without context. That is, subjects had no domain information to help them understand the problem. Alty and Rigas [5] identified the importance of context in enabling blind users to make use of a musical diagram reader. They identified three levels of perception that are important when using auditory interfaces, viz.: detectable mapping, perceptual context and reasoning levels. These may be rephrased as *uniqueness level*, *metaphorical level* and *semantic level* [153].

Uniqueness (or *detectable musical mapping*) corresponds to the construct auralisation’s ability to be uniquely identified and not confused with another. The metaphorical level (or *perceptual context*) is where, given a detectable mapping, the motif creates expectation of the part of the listener. The motif is interpreted in domain terms and meaning can be assigned to it, although the listener may not necessarily be able to reason about the global interactions. The semantic level (or *reasoning level*) is where the listener develops high-level structures in the mind to understand the domain from a higher, or more abstract viewpoint.

Alty and Rigas observed that the production of unique mappings is necessary but not sufficient for a successful auditory-interface design. What is needed is the contribution of the metaphoric and semantic levels. At the metaphoric level the lis-

tener interprets the audio messages in the context of the other messages. The listener begins to interpret the messages in the context of the domain, thus setting up expectations. It is at the semantic level that listeners begin to form gestalts and start to reason about what the audio messages mean.

Applying this to programming and program auralisation means that the construction of unique motifs provides mappings between the program domain and the auditory domain at the uniqueness level. At this level, little meaning is attached, and all that is heard is a collection of different tunes.

At the metaphoric level, users would recognise the tunes as constructs. Metaphors are built in the mind of the listener allowing him to recognise the rising and falling tune as an `IF` statement, or the pleasing flute melody as a `FOR` loop. Once this level is achieved then expectations arise. If a `WHILE` loop is recognised, then the listener would expect to hear the tunes representing the various points of interest of the `WHILE` statement. That is, once the opening point of interest is heard and the motif recognised as a `WHILE` loop, the expectation is then to hear one or more major or minor chord devices followed by the tune signifying the end of the construct. At the semantic level, one starts attributing meaning to these expectations. If the construct is a `WHILE` statement, that means it is a loop and we would then expect to hear one or more evaluations of the controlling condition and the consequent execution of the statement block when the condition is *True*. Further, within the statement block (and before the closing point of interest is sounded) we would anticipate hearing other statements (either simple statements represented by a single percussive sound, or further constructs with their own motif sequence).

At the semantic level the listener is not simply hearing a set of different and recognisable tunes, or even a collection of constructs, but a program and its various branches and interactions. Alty and Rigas claim that the level of mental activity required to perform at this level is likely to increase the memorability of the interface.

Alty and Rigas found in an experiment using the diagram reader, subjects' accuracy improved greatly if a contextual clue regarding the nature of the picture being described was given. For instance, the outline of the letter E was recognised most by subjects who were told that the diagram was a letter; subjects who were given no clue about the picture scored much worse.

In this experiment, the listening test is, in effect, a test of subjects' short term memory and their ability to discriminate between some subtle and some gross differences in motifs. The best level of perception that might reasonably be achieved by this experiment is the metaphoric level, that is, subjects would start to hear auralisations not as separate tunes, but as constructs. If subjects do not move beyond the uniqueness level, then the test is simply one of memory and the subject's ability to correctly repeatedly assign the various unique tunes to one of eight possibilities.

In a real programming and debugging situation subjects would be listening to auralisations at the same time as having access to the program source (even blind programmers would have some textual/verbal representation of the code). The extra context provided by the presence of other constructs and understanding of the aims and structure of the program would allow users to move beyond the metaphoric level into the semantic level. Therefore, that specific identity scores are not high is not of concern as we would expect the added context of the program source to fill in the gaps. Of course, it would be preferable to specify the auralisations such that they are identifiable regardless of context.

5.6.6.5. Learning effect

There was no evidence of a learning effect either within the first listening test or across the two tests. Given that in the first test the two sets of twenty exercises contained the same twenty auralisations (but in different orders) one might have expected some sort of learning effect to be present. That no effect was in evidence, even between the two tests which were separated by a fifteen minutes break is, at first, puzzling. One might think that with repeated exposure subjects would be better able to recognise the individual auralisation tunes. A possible explanation for this is that the whole experiment, which included the introduction and tutorial sessions, lasted nearly three hours. It is quite possible that fatigue played a role in subjects failing to improve their scores over time.

5.6.6.6. The influence of guesswork

We have argued that guesswork did not overly affect the results obtained. Where there is evidence of guesswork is in the disparity between mistaken identification of iterations as selections and vice versa. Recall that no selections were misidentified as iterations whilst 5% of iterations were incorrectly identified as selec-

tions. This would suggest that, when in doubt, subjects tended to assume that the construct was a selection and that the extra clue given by the addition of the auditory parentheses was ignored by subjects when they were very unsure of the identity of a construct. There is evidence of this practice. Subject 7 wrote (see appendix C):

“... The loops, especially nested, were easier to spot and so anything that wasn't I assumed to be a selection”.

Table 5.10 summarises the misidentified iterations. We observe that the **WHILE** construct accounted for the most incorrect identifications. The construct type that was most chosen for incorrectly identified **WHILE** loops was the sub-class *IFs*. Note, of these twenty-two incorrect responses, ten were the result of subject 19.

Iteration	Selec.					Total
WHILE	1	4	1	2	-	8
REPEAT	-	2	-	1	-	3
FOR...TO	2	1	1	1	-	5
FOR...DOWNTO	1	1	3	-	1	6
Totals					1	22

Table 5.10 Misidentification of iterations: listening test 1

The table shows how the various iteration construct motifs were misidentified in the experiment. Each row holds the misidentification results for an iteration construct. We see that the *WHILE* loop was identified once as a generic selection, four times as the sub-class *IFs*, once as a simple *IF* statement, twice as an *IF...ELSE* construct, and never as a *CASE*.

5.6.6.7. Differences between construct types

The two charts below (Chart 5.6 & Chart 5.7) show the *specific identity*, *sub-class* and *class* scores for the forty auralisations in the first listening test.

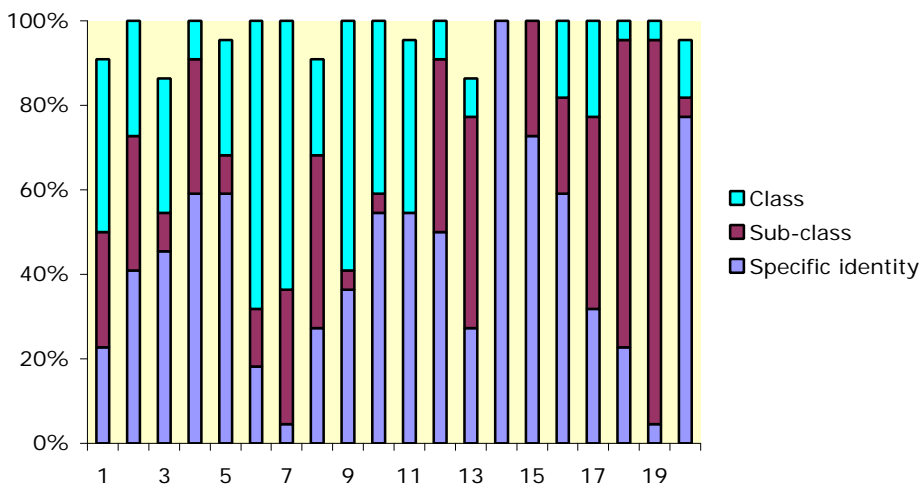


Chart 5.6 Test1 : scores by auralisations 1-20

The chart shows how each of the first twenty auralisations in the listening test were correctly identified at the three levels of identification. For instance, we see that construct 14 was correctly identified by all subjects at its specific identity level; construct 19 was rarely identified at the specific identity level, but by many subjects at the sub-class level.

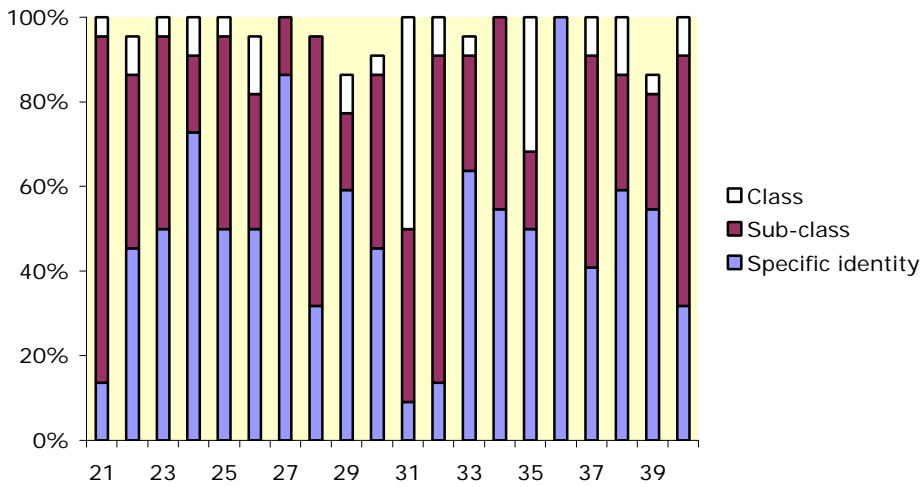


Chart 5.7 Test1 : scores by auralisations 21-40

The scores for the remaining twenty construct auralisations are given in this table.

Some constructs were better identified than others. For instance, constructs 7, 19, 21 and 31 all had very low specific identity scores as compared to constructs 14 and 36 which had perfect scores. From Table 5.3 we see that constructs 7, 19, 21 and 31 were all CASE...ELSEs whilst 14 and 36 were both FOR...TO loops. A summary of identification rates by construct type is given below as Chart 5.8.

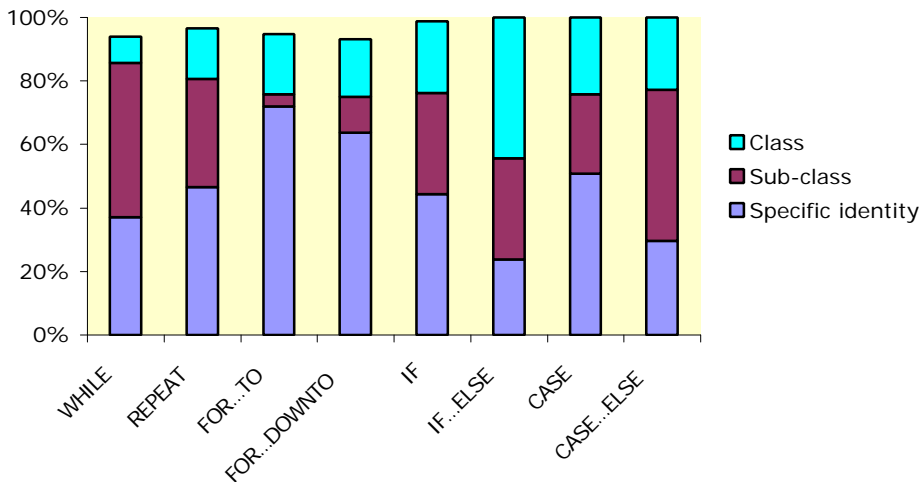


Chart 5.8 Test1: Scores by construct

The identification scores broken down by construct type are given in this chart. It shows that the IF...ELSE and CASE...ELSE were the least successfully identified whilst the FOR...TO and the FOR...DOWNTO had the highest specific identity identification rates.

IF...ELSE and CASE...ELSE were worst performers, tending to be identified as their siblings or at the sub-class level. Most successfully recognised was the FOR...TO, closely followed by the FOR...DOWNTO. Perhaps this is because their motif is more tune-

ful and melodic; also, the drone forms part of the melody. Possibly an additional cue needs to be put into the motifs at the beginning to signify when an ELSE path is present.

5.6.6.8. Subject feedback

In their written responses to the experiment (see appendix C), insights are found into how the subjects perceived the auralisations. The role of timbre was seen as particularly important and six subjects (2, 4, 6, 9, 18, and 20) commented on this explicitly.

Another common comment was that some of the motifs sounded too similar (subjects 3, 5, 11, 13, and 20). Subject 13 commented that the REPEAT and WHILE loops sounded too similar. Conversely, subject 4 found there to be no problem distinguishing between these two loops.

Several subjects felt that familiarity with the motifs would make them easier to recognise. Subjects 21 and 22 both said that they were starting to “get the hang of it” towards the end of the session.

Subject 11 (who scored 13 on the musical knowledge test) gave a particularly insightful summary:

“The tunes weren’t different enough. Most being variations on runs within a scale making it difficult to distinguish. It would have been better to have the tunes varying in tempo as well as melody. By making each significantly different in terms of voice, tempo and tune it would have been easier. Use of rhythm on the selection/iteration made it easy to spot, this could have been used further down the tree.”

This subject’s comments tend to confirm the organising principles put forth in section 5.4.

5.6.7. Conclusions

Following the construction of a program auralisation system, experimentation was carried out to determine the ability of programmers to recognise constructs from their musical auralisations alone. These studies showed that using the auralisations alone (as presently specified) was sufficient for a significant level of accuracy. When the guidelines put forth in the organising principles (section 5.4) are not adhered to, confusion on the part of the subjects arises with constructs frequently

being mistaken for their siblings or subjects unable to differentiate between siblings and choosing a sub-class description instead.

More generally, we know that subtle differences, when augmented by contextual clues and adequate training can be discriminated. For example, parents of identical twins can usually tell them apart. But, ask parents to make absolute judgements (that is, show them only one of the twins) and they are more prone to error. Show the pair of twins to a stranger and they will have difficulty telling them apart even when they are standing side-by-side. In human speech there are about eight or ten dimensions (called *distinctive features*, c.f. CAITLIN's *points-of-interest*) that distinguish one phoneme from another [118].

In programming, we are not really trying to communicate information about a construct's place on a taxonomy chart. Programmers are not interested in these details (just as listeners are not intrinsically interested in the hierarchical structures of music— they simply use those structures to organise the music cognitively). The programmer will already know from the source code listing what constructs have been used and whether or not the various selections have ELSE branches or not. For the purposes of programming and debugging, the auralisations simply have to be good enough to be recognisable within a given context. What is of interest is the structure of the program under consideration, not the categorisation of the various language constructs. Of course, such a hierarchical design approach is not without its merits, as it should allow a programmer to listen to a program auralisation at varying levels of abstraction. To get a feel for program flow it may be enough simply to hear that there is some form of selection here, some form of unbounded loop there; exact details can be gleaned from the listing. Of course, with training and continued use, it may be that users become adept at distinguishing between all the construct auralisations even without contextual clues. This is certainly possible at some level as the designer of the motifs can recognise the individual constructs by ear.

The next chapter describes a study that aimed to assess the usefulness of the auralisation motifs in assisting advanced-novice programmers to locate bugs in short Pascal programs.