A Computer-Assisted Method for Training and Researching Timbre Memory and Evaluation Skills

René Quesnel

Faculty of Music, McGill University

December 2001

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements of the degree of Doctor of Philosophy

© René Quesnel 2001
The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L’auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L’auteur conserve la propriété du droit d’auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.
But what need is there to speak at length concerning the error of the senses, when this same faculty of perceiving is neither equal in all persons nor equal in the same person at all times? Anyone who aspires to search for truth would to no purpose trust wavering judgment.

Boethius (d. 524)

*De institutione musica*
Abstract

Timbre is a multidimensional attribute of sound and depends to a large extent on its spectral content. The evaluation and control of timbre is a task commonly performed by sound engineers, loudspeaker designers, and subjects participating in listening tests on the quality of transmitted and reproduced sound. Such listening tasks require specific listening abilities.

This dissertation presents a training method that aims at developing memory for timbre, sensitivity to timbre changes, and listening strategies involving disciplined auditory attention and efficiency. The physical timbre space is divided into categories defined by the center frequency of standard octave and third-octave resonances. This simplification of the physical timbre space allows the memorization of a limited set of perceptual timbre categories or references that can be used to evaluate other timbres.

The proposed method combines the use of computer software for the presentation and evaluation of exercises and individual tutoring sessions with an instructor. The software monitors the actions of listeners during equalization adjustment tasks and allows data to be collected on the sequence of problem solving operators that are used. Performance indicators can thus be compiled and used for the assessment of skill level achieved by the listeners. A global performance index is proposed.

A listening test involving the adjustment of multiple peaks and dips in a sound's spectrum is presented, in which the performance of a group of experienced professionals
Abstract

involved in sound recording and audio post-production was compared with a group of student subjects with comparatively little listening experience but who were trained with the method described in the present work for a period of 6 months. Results indicate that overall the experimental student group outperformed the control group of professionals. Suggestions for further improvement of the training method and performance level assessment are proposed.
Résumé

Le timbre est un attribut subjectif du son et dépend dans une large mesure du contenu spectral de celui-ci. L'évaluation et le contrôle du timbre sonore font partie des tâches courantes des ingénieurs du son, des concepteurs de haut-parleurs et des sujets participant à des tests d'évaluation de la qualité de transmission et de reproduction électro-acoustique. De telles tâches nécessitent de la part d'un auditeur des habiletés d'écoute spécifiques.

Cette thèse présente une méthode de formation visant à développer la mémoire à long terme du timbre, à accroître la sensibilité aux variations de timbre, et à l'apprentissage de stratégies efficaces d'écoute impliquant une attention auditive disciplinée. L'espace physique du timbre est divisé en catégories qui correspondent à des résonances spectrales centrées autour des fréquences standard d'octaves et de tiers d'octaves. Cette économie de l'espace physique du timbre permet en retour une simplification de l'espace perceptif, facilitant ainsi la mémorisation d'un nombre limité de références pouvant servir à l'évaluation d'un grand nombre de timbres.

La méthode proposée combine l'utilisation d'un logiciel pour la présentation et l'évaluation d'exercices et des séances individuelles de tutorat avec un instructeur. Un mécanisme d'observation des auditeurs pendant la résolution de problèmes permet au logiciel de faire la collecte de données décrivant la séquence des opérateurs utilisés pour atteindre une solution. Ces données permettent la compilation d'indicateurs de
Résumé

performance pouvant être utilisés pour évaluer le niveau d'habileté d'un auditeur. Un indice global de niveau de performance est proposé.

Un test d'écoute nécessitant des ajustements multiples du spectre d'un son est présenté. La performance d'un groupe de sujets ayant suivi la formation pendant six mois y a été comparée à celle de professionnels expérimentés œuvrant dans le domaine de l'enregistrement sonore et de la post-production audio et n'ayant pas suivi un tel entraînement. Les résultats obtenus indiquent que les sujets ayant suivi la formation ont, globalement, surpassé les sujets expérimentés. Les résultats suggèrent des pistes intéressantes de recherche pour améliorer la formation et la méthode d'évaluation de la performance.
Acknowledgements

My first and foremost thanks go to Prof. Donald McLean, Dean of the Music Faculty, who believed in this thesis and in my ability to complete it; and did everything he could to make sure I believed it too. I would like to thank my two thesis supervisors: Dr. Wiesław Woszczyk for supporting this project when it was still a wild dream in my mind, a long, long time ago, for his patience, and his guidance; Dr. Daniel Levitin who jumped into this adventure with enthusiasm, competence, and the sincere, warm support I needed. I would like to thank all my students who took the Technical Ear Training course over the years and gave me invaluable feedback and helped make all this meaningful. I learned a lot from them. I would like to thank the audio professionals who were humble enough to have their listening abilities evaluated by participating in the listening test. Special thanks go to Villy Hansen, Poul Praestgaard, Jens Rahbek and Dr. Søren Bech from Bang & Olufsen A/S, Denmark, for their interest in my work.

J'aimerais enfin remercier Lucie pour les lumières; Gilles et Georgette pour leur support indéfectible et leur confiance inébranlable; Denis, Monique, et Roger, pour tout le reste.
# Table of Contents

Abstract .......................................................................................................................... ii  
Résumé ........................................................................................................................... iv  
Acknowledgements .......................................................................................................... vi  
Table of Contents ........................................................................................................... vii  
Chapter 1: Introduction .................................................................................................... 1  
  1.1 Finding one’s way through timbre space ................................................................. 1  
  1.2 Training of timbre evaluation listening skills ......................................................... 2  
    1.2.1 Auditory perceptual learning ........................................................................ 2  
  1.3 Applications ........................................................................................................... 6  
  1.4 Scope of the thesis ................................................................................................. 7  
Chapter 2: Background .................................................................................................... 8  
  2.1 Introduction ........................................................................................................... 8  
  2.2 Categorizing timbre .............................................................................................. 10  
    2.2.1 Perceptual dimensions .................................................................................. 10  
    2.2.2 Spectral centroids ...................................................................................... 11  
    2.2.3 Vocabularies for timbre description .............................................................. 12  
    2.2.4 Frequency ranges ....................................................................................... 13  
    2.2.5 Formant-based categorization .................................................................... 14
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.2.6 Categorizing for learning</td>
<td>17</td>
</tr>
<tr>
<td>2.3 Training methods for the development of timbre evaluation skills</td>
<td>18</td>
</tr>
<tr>
<td>2.3.1 Timbre Solfeggio</td>
<td>19</td>
</tr>
<tr>
<td>2.3.2 Timbral Ear Trainer 1</td>
<td>20</td>
</tr>
<tr>
<td>2.3.3 Spectral Ear Training</td>
<td>21</td>
</tr>
<tr>
<td>2.3.4 Training subjects for listening tests</td>
<td>22</td>
</tr>
<tr>
<td>2.3.5 Other training programs</td>
<td>24</td>
</tr>
<tr>
<td>2.3.6 Limitations of previous methods</td>
<td>25</td>
</tr>
<tr>
<td>Chapter 3: Training Methods and Issues</td>
<td>29</td>
</tr>
<tr>
<td>3.1 Introduction</td>
<td>29</td>
</tr>
<tr>
<td>3.2 Adaptive training methods</td>
<td>30</td>
</tr>
<tr>
<td>3.2.1 Extended drill-and-practice</td>
<td>30</td>
</tr>
<tr>
<td>3.2.2 Other adaptive methods</td>
<td>31</td>
</tr>
<tr>
<td>3.3 A monitoring system</td>
<td>33</td>
</tr>
<tr>
<td>3.4 Evaluation criteria</td>
<td>34</td>
</tr>
<tr>
<td>3.5 Deriving a performance index</td>
<td>37</td>
</tr>
<tr>
<td>3.6 How to cope with erroneous data</td>
<td>38</td>
</tr>
<tr>
<td>3.6.1 Listening fatigue</td>
<td>38</td>
</tr>
<tr>
<td>3.6.2 Lack of attention</td>
<td>39</td>
</tr>
<tr>
<td>3.6.3 Human-computer interaction problems</td>
<td>40</td>
</tr>
</tbody>
</table>
Table of Contents

3.6.4 Manipulation errors .................................................. 41
3.7 How to cope with uncertain perceptibility in contextual listening ................. 42

Chapter 4: Implementation ...................................................................... 44

4.1 Training program overview ....................................................... 44
  4.1.1 Development of memory for timbre ..................................... 45
  4.1.2 Equalization adjustment tasks ........................................... 46
  4.1.3 Perception of dips .......................................................... 46
  4.1.4 Efficiency ................................................................. 47
  4.1.5 Exercise editor ............................................................ 47
  4.1.6 Sound library ............................................................. 48

4.2 Listening task ............................................................................ 49
4.3 Audio tools .............................................................................. 52
4.4 Monitoring architecture overview .............................................. 53

4.5 Observing listeners .................................................................... 56
  4.5.1 Tracing a user's actions and strategies ................................. 56
  4.5.2 Problem solving operators ............................................... 61
  4.5.3 Event trace history ......................................................... 63
  4.5.4 A listener's profile .......................................................... 66

Chapter 5: Validation .......................................................................... 69

5.1 Introduction ................................................................................ 69
# Table of Contents

5.2 Test subjects .................................................................................................................. 70
5.3 Stimulus .......................................................................................................................... 72
5.4 Procedure ....................................................................................................................... 73
  5.4.1 Task ........................................................................................................................... 73
  5.4.2 Score evaluation method ............................................................................................ 77
5.5 Results and discussion .................................................................................................... 78
  5.5.1 Score .......................................................................................................................... 81
  5.5.2 Response time ............................................................................................................ 85
  5.5.3 Listening and usage times .......................................................................................... 86
  5.5.4 Number of steps ........................................................................................................ 87
  5.5.5 Sound comparisons .................................................................................................... 89
  5.5.6 Vowels ....................................................................................................................... 91
  5.5.7 Gain and frequency adjustments .............................................................................. 92
  5.5.8 Listening level ........................................................................................................... 93
  5.5.9 Performance index ..................................................................................................... 94
5.6 Summary ......................................................................................................................... 96

Chapter 6: Conclusions ....................................................................................................... 98
6.1 Summary ......................................................................................................................... 98
  6.1.1 Training program ....................................................................................................... 98
  6.1.2 Measurement of performance level ....................................................................... 99
# Table of Contents

6.1.3 Monitoring problem solving in timbre/equalization adjustment tasks ....... 101

6.2 Limitations and suggestions for further research ............................................. 102

6.2.1 Assessing expertise .................................................................................. 102

6.2.2 Problem solving monitoring ..................................................................... 105

6.2.3 Training Program ...................................................................................... 107

Appendix 1 – Listening Test Protocol ................................................................... 109

Appendix 2 – Development Tools ......................................................................... 114

References ............................................................................................................ 115
Chapter 1: Introduction

1.1 Finding one’s way through timbre space

From Helmholtz (1863) to Hajda, Kendall, Carterette, & Harshberger (1997), there have been numerous attempts to define what *timbre* is (Sandell 1997). A definition of timbre remains somewhat elusive; it has been described as the perceptual property of sound associated with its "quality", as those aspects which serve to distinguish two sounds other than differences in pitch, loudness, (perceived) duration and (perceived) location, but being affected to various degrees by several physical, acoustic aspects of sound. In recent years, timbre research has shown that the spectral content and the fluctuations of the spectrum over time in an acoustic signal are major contributors to a sound’s timbre (e.g., Krumhansl 1989; McAdams, Winsberg, Donnadieu, Soete, & Krimphoff 1995).

Individuals who must, for their work, rely on their ability to identify, discriminate and evaluate sounds based on their timbre are faced with a daunting task, for which the required skills are most often acquired slowly over a long period of time. Such professionals include acousticians designing concert halls, engineers designing loudspeakers and hearing aids, telecommunication engineers, sonar operators, subjects
participating in listening tests, and recording engineers and producers. All share a common need: to find their way through timbre space in order to recognize, identify, evaluate, name, or control distinctive and salient timbral features of recorded and transmitted sound, in concert halls, or from audio equipment.

"Timbre space" should be understood here as an n-dimensional perceptual continuum. To be manageable for training and control purposes, this timbre space must be simplified, segmented, partitioned, quantized or otherwise categorized in a limited number of elements, dimensions, categories, or anchor points.

1.2 Training of timbre evaluation listening skills

1.2.1 Auditory perceptual learning

Several studies have demonstrated and documented learning processes of various auditory perceptual tasks (a good summary can be found in Watson 1980). The amount of training necessary to achieve asymptotic performance varies depending on the type of task, the complexity of the signals, the number of different stimuli that are presented and stimulus uncertainty. Typical listening tasks include the detection of a target sine tone in the presence of noise, discrimination of stimuli differing in loudness, pitch, duration, etc.,
Chapter 1: Introduction

and the identification of stimulus parameters such as pitch. Detection, discrimination, and identification require an increasing amount of practice, from a few hours to hundreds of hours. It is typical that differences between listeners are large at the beginning of a training period and diminish significantly at the end. Initial performance is higher and the learning period shorter for sounds that are familiar to the listener. These last two points have been verified by Quesnel (1991) for listening tasks involving discrimination and identification of formant-type spectral manipulations. In recent years, auditory perceptual training of various skills has been recognized as an important factor in listening experiments involving the evaluation of reproduction quality of audio equipment (Bech 1993). It is generally recognized that trained listeners are more reliable, consistent in their judgments, and provide more robust data than naive listeners.

1.2.2 Perceptual learning of timbre

Learning effects in timbre perception have been studied in two areas of research: auditory profile analysis and in the development of courses or training programs for sound engineers and subjects of listening tests.

In auditory profile analysis, subjects are asked to detect differences in the level of one of the components of a multi-component complex signal. The difference is perceived as a change in timbre, as it should be expected since the sound manipulation results in a
change of spectral shape. The profile analysis theory states that listeners perform simultaneous comparisons of different parts of the spectrum of a sound stimulus instead of successive comparisons of individual bands; this implies that the comparisons constitute a parallel cognitive process rather than a serial one (e.g. Anderson 1990). It has been observed that thresholds are lower for wide-band signals. Typically, sound stimuli in profile analysis experiments are constructed by assembling sinusoidal components spaced logarithmically between say, 500 Hz and 3 kHz. The change in spectrum is accomplished by adding for example a 1 kHz component in phase with the 1 kHz component already present in the reference or standard stimulus. Learning effects in profile analysis tasks have been investigated by Bech (1991), Kidd, Mason, & Green (1986), and more recently by Drennan & Watson (2001). Bech (1991) found out that learning rate varied greatly between a group of 10 listeners over 1200 trials. Some listeners improved significantly over the training period while others were resistant to the training. The experimental setup didn't allow any conclusion on this discrepancy between subjects regarding the training effectiveness. Drennan and Watson concluded that while the ability of listeners to discriminate differences in spectral shapes improved with practice, both the learning rate and asymptotic levels of performance that were achieved varied greatly between subjects.

A few training programs and courses have been reported in the literature that aim at developing the ability to recognize, discriminate, evaluate and identify spectral resonances and anti-resonances (Brixen 1993; Letowski 1985; Miskiewicz 1992; Olive
Chapter 1: Introduction

1994 & 2001; Quesnel 1991 & 1996). Target populations for these programs are typically audio engineers and subjects participating in listening tests for the evaluation of consumer and professional audio equipment, in the context of research and development by manufacturers and designers of equipment. There is also clear relevance to military and other applications (e.g. sonar, signal detection) although these connections are not drawn explicitly in these courses. All programs have been reported to improve the listening skills of their users. However, much is still unknown about the processes involved in the learning of these skills, what are the problem solving strategies an expert listener will use when faced, for example, with a complex timbre matching task. Complex listening tasks can involve a set of perceptual and cognitive skills that, at different levels of perception, will contribute to the successful evaluation of timbre differences. These skills have not yet been clearly identified and systematically investigated.

The issue of determining how to evaluate the skill level of a listener following a training program has been addressed in various ways but I believe none of them are completely satisfactory. This important issue of determining the expertise level of a listener is among the primary topics examined in this thesis.
Chapter 1: Introduction

1.3 Applications

The process of adjusting spectral characteristics of sound in the course of professional tasks such as recording and mixing by a recording engineer, adjusting timbral parameters of a synthesized sound by a composer or sound designer, adjusting subjectively the frequency response of a loudspeaker or hearing aid by an engineer, is poorly if at all documented. Although there exist many books about how to place microphones in order to achieve particular desired effects, the listening strategies necessary for obtaining and evaluating these effects are rarely discussed; in general, there is little published research on the manipulation of timbre towards producing a given sound goal and how the internal cognitive processes of experts – largely operating on intuition – might be "externalized" for subsequent teaching. The research presented here has potentially two broad categories of applications: First, the study of expert and highly skilled audio practitioners (sound engineers, loudspeaker designers, professional listeners, composers, sonar operators) may provide invaluable insights into the processes, methods, and strategies that are successful when the spectral content and hence the timbre of sounds need to be manipulated to achieve a given goal. This information can provide a substantial contribution to the knowledge of perception and evaluation of timbre in the field, in varying and complex auditory contexts and for various purposes. Second, an efficient method of developing the required skills for reliable evaluation of timbre and
timbre differences can potentially benefit a large number of professionals involved in the evaluation, manipulation, and control of sound quality in general and of the timbral aspects of sound quality in particular.

1.4 Scope of the thesis

This dissertation focuses on a new method developed to observe listeners as they work through the solution of a spectral adjustment task. An important goal of this is to better understand the actions a listener is performing during timbral manipulation tasks and to subsequently use this information to improve the usefulness and efficiency of the training. The training method has been implemented in computer software that allows listeners to train and data to be gathered, in combination with individual tutoring sessions with a human instructor.

The task used to illustrate and validate the training method requires the listener to cancel spectral modifications that were applied by the training software to a given sound stimulus. The task is sufficiently complex to demonstrate several sub-skills that are believed to be instrumental in performing this type of task efficiently. The task has the advantage of sharing similarities with a common task performed in sound recording practice, sound synthesis, and loudspeaker design (among others): to remove undesirable spectral resonances and to compensate for unwanted dips in a sound's spectrum.
Chapter 2: Background

2.1 Introduction

A large body of work exists on the perception of timbre and the recognition of sound sources. A significant amount of work has also been published on the detection of changes in spectral shape of multi-component synthesized signals (profile analysis) but much less has been published on the training of timbre evaluation skills, let alone on the perception of timbre in the context of spectral adjustments performed on recorded and transmitted sounds. Much of the efforts deployed in the area of timbre research have been devoted to ways of simplifying the constructs of timbre to make it more manageable for both study and experimentation and for synthesis purposes.

It is well established that timbre is a multidimensional attribute of sound (Grey 1975; Krumhansl 1989; McAdams et al. 1995), both in terms of its physical and psychological descriptions. This multidimensionality of timbre complicates greatly its study as it is difficult to isolate and even more assess the relative importance of the numerous physical parameters of sound that contribute to its perceptual variations. The non-linearity of the ear's transfer function (Moore 1989) contributes further to the complexity of the task.
Chapter 2: Background

Section 2.2 in this chapter looks at previous timbre research from the perspective of the need to simplify the psychological timbre space as it is an important requirement for the training of timbre listening skills. It is important to note also that to be precise, timbre is not viewed here as simply that which differentiates say, an oboe from a clarinet sound (Saldanha and Corso 1964). These two instruments certainly have different timbres. But at a finer grain of analysis, it also becomes apparent that different clarinet players can have distinctive timbres, the same clarinet player can produce sounds with different timbres depending on the carving of the mouthpiece or if the reed that is used is thick or thin, and the different registers of a single clarinet have different timbres (each labeled with its own name). From this dissertation's perspective, two sounds are considered to have a different timbre if the differences in their spectral contents are perceivable, even if only slightly. It is conversely true that two sounds displaying different spectra from an FFT analysis may be perceived as having the same timbre. For example, some measured differences in spectra may be masked by other spectral components. Section 2.3 reviews other methods of training timbre evaluation skills. Finally, a summary of the limitations of these methods is presented.
Chapter 2: Background

2.2 Categorizing timbre

"Categorization" here is taken in its general meaning of subdividing a concept, or perceptual phenomenon into categories, classes that, without necessarily being able to account for all possible variations of the concept or phenomenon, provide nevertheless useful anchor points, references, or subtypes that can be more easily remembered and conceptualized. Categorization can facilitate the conceptualization of a complex perceptual phenomenon (Anderson 1990).

2.2.1 Perceptual dimensions

A great deal of effort in timbre research has focused on the definition or identification of a limited set of perceptual dimensions for timbre, from the pioneering work of John Grey (1975, 1977) to the more recent work by McAdams et al. (1995) and Hajda et al. (1997). Typically, listeners are asked to rate the perceptual similarity or dissimilarity between pairs of a set of synthesized timbres otherwise equal in pitch, loudness, and duration. Multidimensional scaling techniques (Shepard 1963, 1980) are used to build a geometric representation of the perceptual relationships among the synthesized timbres. The distance between pairs of sounds in the resulting Euclidean
space corresponds to their perceived dissimilarity. The experimenters then derive perceptual dimensions and their physical correlates. Results from all multidimensional studies of timbre describe one dimension related to the brightness (spectral centroid) of sound, and a second dimension is related to the attack quality (rise time). The psychophysical interpretation of the third dimension is often more difficult and seems to vary depending on the group of stimuli and the group of subjects. Krumhansl (1989) associated this third dimension with "spectral flux", the micro-temporal fluctuations of the spectral envelope. McAdams et al. (1995) described it as the spectral fine structure of a sound which correlates with non-temporal spectral irregularity. McAdams added to each sound described in the timbre space a specificity value related to attributes unique to each sound.

2.2.2 Spectral centroids

The spectral centroid is a measure of the distribution of spectral energy in a sound and represents the spectral center of gravity or the mid-point in the distribution of energy along the spectrum of a sound. A collection of sounds can be classified in terms of their spectral centroids.

While listeners might potentially be trained to perceive timbre differences among a set of sound samples based on their spectral centroid differences, this attribute is not
readily amenable to training as the audio processing tools commonly used to manipulate the spectrum of sounds do not modify these sounds in terms of their centroids but instead in terms of patterns of resonances and anti-resonances. In addition, spectral centroids account only for a single perceptual dimension. Stimuli varying in spectral shape but maintaining constant centroids can still be discriminated on the basis of their timbre (Kendall & Carterette 1996).

2.2.3 Vocabularies for timbre description

The communication of timbre differences and timbre characteristics is an important issue that presents several difficulties. Words, qualifiers, and descriptors can have different meanings for different people speaking the same language and descriptive words often do not translate well to other languages. There are many more differences in timbre than there are applicable verbal descriptors. Differences in meaning are commonly found among expert listeners, even among a small group of listeners participating in a common training program for the development of their timbre perception skills. Many authors have used sets of descriptive words in studies on the evaluation of sound quality (Faure 2000; Gabrielsson and Sjogren 1979; Letowski 1995; Salmon 1950a, 1950b; Solomon 1958; von Bismarck 1974b). At best, a vocabulary for timbre can be developed and adopted by small groups of listeners for specific purposes (Hansen 1987). But such
vocabularies may not be applicable outside the group. Informal experience I gained by teaching timbral ear training tends to indicate that it is not simple and easy to "impose" a meaning for a word different than the meaning that has been associated for a long time by a given listener. Even if the listener agrees to change the meaning of the word, it will not be used if it is different enough from the meaning previously attached to it by the listener. Descriptive words are certainly not useful in describing very fine differences in timbre. Non-verbal elicitation seems preferable. Rumsey (2000) provides a good overview of the issues related to the use of verbal descriptors for sound quality. Bech (1999) borrows from the field of descriptive analysis to discuss aspects of the development of a descriptive language as related to the training of listening panel subjects. In summary, verbal descriptors for describing timbre present limitations in accuracy, portability, and standardization.

2.2.4 Frequency ranges

The spectrum of a sound can be divided into sub-ranges within the limits of human hearing abilities (roughly between 20 and 20,000 Hz) (Salmon 1950a, Salmon 1950b). Such a division is commonly encountered in consumer audio equipment, such as when a home or car stereo has "bass" and "treble" controls; these controls allow for the manipulation of the spectral content of the sound at two fixed frequencies. More
elaborate controls are incorporated into professional audio devices which manipulate the spectrum more precisely. For example, parametric equalizers (Massenburg 1972) typically divide the audible frequency range into 3 main categories (low, middle, and high frequencies) with further "refinements" such as "low-mid", "high-mid", etc. Unfortunately, different mixing consoles use different subdivisions, and those that have the same nominal subdivisions are often not calibrated to one another. In the Technical Ear training course I teach at McGill University, a simple set of three sub-divisions has been adopted, each subdivision being represented by a subset of frequency values in Hz:

- low frequencies (63, 125, 250 Hz)
- mid-range frequencies (500, 1000, 2000 Hz) and
- high frequencies (4, 8, and 16 kHz).

These ranges are readily accepted by the students and are useful as a first, coarse, subdivision of the whole spectrum.

2.2.5 Formant-based categorization

The concept of "formant" comes from the study of speech where it represents spectral resonances produced by the speech-producing human apparatus and which form
the building blocks of vowels. Flanagan (1956) provided three definitions of formants, one of which allows the concept to be applied outside of speech: Formant refers to the "frequency of the centroid of a gross concentration of spectral energy" (p. 110). Formants have thus been used to describe resonances in the spectrum of musical instruments that remain constant across registers and which provide these instruments with characteristic signatures (Luce 1963; Risset & Wessel 1991). The subdivision of the frequency spectrum perceptual continuum into formant categories has been found to be very useful for the memorization of a set of timbre references (Letowski 1985).

The relationship between timbre of non-speech sounds and vowels has been demonstrated by Slawson (1968) and has been used by others (Letowski 1985; Meyer 1978; Quesnel 1991). According to Slawson, "The vowels serve for sound colors the same function that note names or solfège syllables serve for pitch" (p. 154). In training, the use of "artificial" formants within the vocal range can be associated with vowel-like sounds that provide the listener with a familiar frame of reference facilitating the memorization and subsequent identification of other timbres having similar spectral characteristics. Most listeners initially use local cues in recognizing certain spectral modifications, i.e. listeners attend to the effect a given formant will have on a specific sound production aspect of a given musical instrument. Such effects can be more or less revealed depending on the particular sound recording used and a listener will often need different sets of cues for different types of sound sources. The association of vowels and
formant references has the significant advantage of being applicable to a very large range of audio sources. The long-term memory trace of the timbre of a 1 kHz resonance associated with the "a" vowel that a trained listener acquires can be recognized readily whether the sound source is a solo clarinet, a rock band, a string quartet or pink noise. The timbre references thus acquired are transferable to a large number of types of sound samples. Rakowski (1988) enumerates three criteria for the selection of a set of codes at the phonological level that should be met in order to be effective for auditory communication:

- The units or categories must be easily distinguishable from each other
- The set should not exceed our memory capacity and
- The set should allow the formation of additional sub-categories.

The use of vowels has a number of limitations though.

- Vowels are usually built out of 2 to 3 formants. The vowel references discussed here are characterized by only one formant. As a result, the vowel character can be more or less perceptually distant from a real vowel, depending on the sound sample to which it is applied.
Chapter 2: Background

- Vowels as used here thus can be considered as overlapping categories, the degree of overlap depending on the sound sample being heard, the characteristics of the resonance producing them, and sequencing effects. Therefore, a stimulus (i.e., a sound sample to which a resonance was applied) can sometimes be perceived as being a member of one category and sometimes a member of another category (Ashby & Maddox 1998, p. 253). Many times though, the association of a vowel to a given resonance is very obvious (Peterson & Barney 1952).

- Vowels as formant-based timbre references are relevant within a limited frequency range only, roughly between 250 and 4000 Hz.

- There are not enough vowels to account for all the resonances to be evaluated within the vocal range. A more complete system could possibly be built out of vowels borrowed from different languages but foreign vowels are often difficult to reproduce by a non-native speaker.

2.2.6 Categorizing for learning

Categorization and classification are central aspects of all learning processes (Estes 1994; Harnad 1987). We remember events (or stimuli) and can identify recurrences of them as instances of classes or categories (Estes 1994, p. 4). Timbre space is much too large to be tackled efficiently without some means of simplifying and reducing its
perceptual continuum into more manageable "slices" or "chunks". A reduced set of categories provides the learner with distinct references that can be discriminated among each other. Once an initial limited set of categories has been stored in long term memory, the set can be expanded to include intermediate categories in between each of the initial categories.

2.3 Training methods for the development of timbre evaluation skills

This section reviews existing training programs for the development of timbre evaluation listening skills. The training materials in most of these programs are delivered by computer and are intended for student sound engineers, audio equipment design engineers, and members of listening test panels who evaluate the sound quality of audio equipment. The strengths and limitations of these programs are summarized at the end of the section.
Chapter 2: Background

2.3.1 Timbre Solfeggio

Rakowsky (1979), Rakowski and Trybula (1975), Letowski (1985), Miskiewicz (1992), and Miskiewicz and Letowski (1999) have described a course developed over several years at the Chopin Academy of Music in Warsaw for the development of listening skills associated with the evaluation of timbre and sound quality. The three-year course covers a wide range of audio quality aspects including loudness, masking, distortion, and spatial hearing. The first year is devoted to the timbral aspects with the objectives of developing sensitivity to timbre differences and building long-term memory for a set of timbre references. Timbre changes are produced in two ways: by low- and high-pass filtering and by applying peaks in the spectrum ("formant enhancing"). Both active and passive tasks are used. In passive tasks, students are asked to evaluate the spectrum of sounds without direct manipulation of the sound. An instructor presents various spectral modifications and students provide their answers on prepared answer sheets. In active tasks, students have to match a spectral modification applied by the instructor using either a parametric equalizer or low-, high-, and band-pass filters. At the beginning of the course, a limited set of octave standards are used. Other center frequencies comprising up to 27 1/3-octave standards are then introduced progressively. Unfortunately, only overviews of this seminal work have been published and details about the training methods and resulting data remain unpublished.
2.3.2 Timbral Ear Trainer 1

Based in part on the work of Letowski and Miskiewicz reported above, a computer software application was developed by Quesnel (1991) to implement simple timbre matching and identification tasks. The computer was used to present exercises to the listener, collect the listener's responses and perform a simple evaluation (right/wrong). MIDI-based external parametric equalizers were used to apply resonances and anti-resonances to various sound material: speech, noise, single musical instruments, and ensembles. Three types of exercises were implemented. In the first type ("EQ Matching"), the user had to match the equalization applied by the software by comparing the question (the standard) with a "flat" version of the same sound sample and apply the necessary spectral adjustments until both standard and the working samples had the same timbre. In a second type of exercise, the listener's task was not to match but to cancel the equalization applied on the standard by applying a mirror equalization to the sound sample. Finally, the third type of task was to identify the center frequencies of resonances. This exercise relied on the memorization in long-term memory of timbre categories corresponding to formants based on octave frequencies from 63 Hz to 16 kHz.

A pre-test/post-test procedure was used to compare the performance of 8 students before and after the training relative to three aspects: 1) score (percentage of correct answers) on an EQ matching task; 2) mean response time on the same task, and 3) score
on an absolute identification task. While response time data was recorded, the students' scores were calculated in terms of response accuracy only. Results from the EQ matching task indicated two main trends. First, all students' scores improved and the amount of improvement varied much between students. Second, the training had the effect of reducing the performance variability among the listeners who all but one obtained perfect scores at the end of the training.

2.3.3 Spectral Ear Training

Brixen (1993) presented a training program for audio engineers working at the Danish Acoustical Institute to develop their ability to recognize octave and 1/3 octave center frequencies of peaks and dips. Frequency values range between 32 and 16 kHz, available gain values are 1, 3, 6, 9, and 12 dB and the Q value is fixed. Up to two bands can be modified simultaneously and peaks and dips can be mixed. The training has a duration of 2 days with an instructor with a period of 3 to 4 weeks of self-training in between. The training with the instructor is computer-based but the self training is carried out by listening to pre-recorded tapes. No results are provided but the author concludes that it is possible to some extent to become a human spectrum analyzer within a short training period.
Auditory experiments are often conducted in which subjects are asked to evaluate various qualitative aspects of audio reproduction systems. Many authors have recognized the importance of skilled, experience listeners in such experiments as they tend to provide more reliable and detailed judgments (Bech 1992, 1993; Toole 1985).

Bech (1993) studied training aspects associated with a listening task in which subjects had to detect a timbre difference between a sound and the same sound accompanied by an early reflection. He found that threshold decreased significantly with training and listeners reached asymptotic performance after 1200 trials. The skills developed during the training were found to transfer to a certain extent to other listening tasks of the same nature. Stability of the training was evaluated to be approximately 4 months.

Olive (1994, 1998, 2001) presented the successive versions of a computer-based system to train listeners to detect, identify and rate typical peaks, dips, and bandwidth limitations found in loudspeakers. Two main types of activities are implemented. In identification tasks, the listener has to select among a choice of bands ranging from 2 to 12, which one contains the spectral modification applied by the computer. In the standard training configuration, the Q value is adjusted automatically based on the number of possible bands and ranges from 0.24 to 1.44. It is also possible to set a fixed Q but this is
Chapter 2: Background

not part of the standard training. Center frequencies cannot be user-specified and are set automatically based on the number of possible bands the listener has to choose from. Frequency values vary depending on the number of bands and range between 13.8 Hz to 11536.5 Hz. Default gain setting for peak/dip spectral modifications is ±6 dB. In the default training setup, the listener has to identify a single frequency band but up to three bands can be equalized at the same time. The skill level is determined by the number of bands the listener can choose from and ranges from 2 to 12. For each skill level, a passing score and the minimum number of trials to answer can be specified. In a given run of trials, multiple sound stimuli can be selected. The number of trials for a given level is by default set to the number of selected sound stimuli. The second type of training task simulates an actual listening test in which the listener rates 4 different stimuli on a preference rating scale ranging from 0 to 10. One of the stimuli is a hidden, "flat" or non-equalized reference. Preference rating training has three levels of difficulty, with the perceptual difference between the stimuli decreasing as the difficulty level increases. Preference tests are used to train listeners with the task of rating products and measure their discrimination skills and consistency. The software includes a basic statistic method to evaluate the reliability of judgment of the listeners, which is central in listening tests. Olive doesn’t report on the typical level achieved by listeners and the time it takes for listeners to achieve level 12. In Olive (1994, 1998), the effect of various parameters and their interactions on the performance of listeners were examined. The choice of sound
sample was found to have the strongest effect on listener performance, with sounds of larger music ensembles and wider spectra being easier to work with than smaller ensembles and solo instruments.

The experience I gained teaching timbral ear training at McGill suggests that a critical factor is the spectral balance of the sound sample. A jazz trio recording with an evenly balanced spectrum may be a better sound material for detecting equalization manipulations than a large orchestral recording with strong resonances tending to mask those introduced as part of the training. Olive reported that dips in the spectrum are commonly confused as peaks at frequencies slightly above the dip. This may be true of naive listeners or with limited experience but results obtained in the timbral ear training course at McGill suggest that trained listeners rarely confuse peaks and dips.

2.3.5 Other training programs

Two commercially available training programs currently exist for the development of technical listening skills, and the target users are professional audio engineers. Everest (1982, 1997) describes a course in technical listening covering several aspects of sound quality (including timbre, non-harmonic distortion, reverberation, etc.). There are 10 lessons in the course. Lessons 1 to 9 present theoretical concepts with audio illustrations. Lesson 10 summarizes the course contents and contains a quiz. Moulton
Chapter 2: Background

(1993) developed an audio ear training course for recording engineers and musicians which cover a wide range of aspects of sound quality such as equalization, delays, reverberation, distortion, loudness, compression, etc. The training related to the perception of peaks and dips in the spectrum contains 2 volumes. The "advanced EQ" section covers 1/3-octave single peaks and dips and double octave peaks and dips. The drills are in the form of A/B comparisons with A being the reference (flat) signal and B the processed signal. The user answers the drills in a workbook.

2.3.6 Limitations of previous methods

This section describes the most important limitations of the training programs and courses described above. Some of these courses and training programs cover several aspects of audio quality. I concentrate here on the timbral aspects only.

The Timbre Solfeggio course is not very efficient in its delivery method. Students meet the instructor once a week. In the active exercises, the instructor applies spectral modifications to a sound sample and students come in turn and attempt to match the equalizations. In passive exercises, students write the description of equalizations applied by the instructor on response sheets and then discuss their answers. Individual practice is not possible.
The format of both Everest's and Moulton's courses precludes any adaptation of the training content to individual needs and skill level and it cannot be expanded. Everest's seems more theoretical in nature with few exercises. The difficulty level of the exercises in the timbre section of the Golden Ears course is not very high.

The extent of the training offered by the Spectral Ear Training program is clearly insufficient. Novice listeners can certainly improve on the basic skills in a relatively short period of time but it seems overly optimistic to state that a listener can become a "human spectrum analyzer" in a few weeks. It of course depends on what being a human spectrum analyzer means.

The training described by Olive is targeted towards member of listening tests panels. The types of spectral modifications that are applied by default are implemented to reproduce resonances and band limitations typically found in loudspeakers. Other frequency and Q values can be specified but they are not part of the standard training. There does not seem to be a systematic effort to develop long-term memory for timbre references. Each level, which corresponds to the number of frequency bands to choose from, has a different set of center frequencies. As a result there are a total of 77, non-standard center frequencies that listeners work with. This not only exceeds human information processing limits established by Miller (1956) but also exceeds the labeling skills of many of the best absolute pitch possessors who are limited to the 12 chromatic tones due to persistent octave errors (Takeuchi & Hulse 1993).
Chapter 2: Background

It seems doubtful that an effective categorization of the timbre space can be achieved this way and facilitate the storing in long-term memory of robust references. It would be interesting to know then what do the listeners learn. The maximum resolution for center frequencies is 0.9 octave. However, depending on the signal and the spectral region, the human hear can easily hear at least 1/3 octaves (0.3) and depending on the signal, the resolution can be higher. Except for the preference test exercises which implement real-world contexts for subjects of listening tests, the matching exercises are in essence multiple choice questions, the number of possible choices defining the skill level. Of course, as the number of choices increases, the perceptual difference between the different bands to choose from decreases, making the task increasingly difficult. When a listener chooses an incorrect frequency band, the feedback consists of displaying a window notifying the listener that the answer was wrong and showing visually what was the right band. But the listener cannot go back and compare aurally the answered settings with the question's. The only choice is to click on the "next" button. This type of interaction and task is not appropriate for sound engineers who have, on a daily basis, to actively modify the spectrum of sound with equalizers that typically allow continuous adjustability of frequency and gain.

The Timbral Ear Training software (version 1) provided very limited feedback to the user regarding his performance. The software performed a strict comparison of the data describing a question and the data describing the answer and had no knowledge
Chapter 2: Background

about the nature of the task, possible confusions, listening strategies that can be useful in solving listening tasks such as Equalization Removing. In addition, measurement of performance level was based on accuracy only. Experience gained by the present author in teaching timbral ear training has shown clearly that response time can become an important factor in assessing the ability of a listener to perform formant-type spectral modifications recognition, matching, and identification.
Chapter 3: Training Methods and Issues

3.1 Introduction

The effectiveness of a training program, whether computer-based or delivered in a classroom can be measured, in part, by the time it takes for learners to achieve a target level of performance and the skill level that the program allows the learners to attain (du Boulay 2000). Effectiveness can also be affected by the level of interest and motivation the program can raise.

As discussed previously, untrained listeners may show a high level of variability in their listening skill levels for many types of auditory tasks. In timbral ear training, not all listeners have the same listening background, experience, and skills at the start of a training program. Not all listeners have the same training needs (e.g. recording engineers, designers of loudspeakers, listening test subjects), learn at the same pace, or react equally well to certain types of training. Not all listeners have the same learning difficulties, the same motivation or attitude, or make the same mistakes for the same reasons. Not all wrong answers indiscriminately ask for the same feedback from the software program. This chapter examines issues involved in the implementation of effective training in a computer program, with an emphasis on the difficulties related to the training of timbre evaluation skills. An overview of various options available to make a computer software
Chapter 3: Training Methods and Issues

program adaptive to some aspects of training is presented. The rationale for the choice of the particular method selected for the present work is presented and discussed.

3.2 Adaptive training methods

"Adaptive training" can be implemented at various levels of complexity and can carry varied meanings. Adaptivity can be implemented as contextual feedback to the user, can involve the generation of questions tailored to the current ability level of the user, and can also involve the design of complete and dynamic training programs tailored to individuals with the ability to present customized remedial actions for various types of errors.

3.2.1 Extended drill-and-practice

To be efficient, a training method must go beyond the “drill-and-practice” approach in which questions are presented in random order regardless of the performance achieved by listeners on specific items. In standard drill-and-practice, a single pool of questions is repeatedly presented to the listener until a target performance level is achieved. Statistically, items that are mastered by the trainee are presented as often as
items for which the subject has more difficulty. While the presentation of mastered items
can reinforce to a certain extent the "learned" state of the item, more difficult items are
not presented as often as would be required in order for the subject to attain a similar
level of performance.

Simple adaptive training methods can extend or refine the standard drill-and-practice approach by identifying mastered items and providing, for example, two pools of questions, a working pool and a review pool (Salisbury 1988). Items that are identified as mastered based on some criterion are transferred to the review pool. On subsequent presentation of the question set, the proportion of non-mastered items (from the working pool) and mastered items (from the review pool) increases until all items are mastered. Other variants are possible.

3.2.2 Other adaptive methods

Some degree of adaptivity can be achieved if different training programs (the contents of the training) can be defined for individuals or groups with different needs and training goals. Others such as Lajoie (1993) have designed learning systems in which everyone receives the same instruction and performs the same tasks but the help and feedback are adjusted to the context and expertise level of each user. Effective training
and feedback can be obtained by implementing in the software a good representation of
the problem to solve and a representation of what the learner is doing (Reusser 1993).

We still know too little about the skills, knowledge, and cognitive processes
involved during complex spectral adjustments to be able to develop accurate, moment-to­
moment modeling of the cognitive state of a real listener in the process of evaluating the
spectral content of a sound sample. Research in machine listening (Martin 1999) has
developed sophisticated models over the past few years. There are now real-time systems
capable of classifying sound samples from several musical instruments played using
different techniques e.g. (Fujinaga, 2000) but these remain coarse recognition systems in
comparisons to spectral subtleties that a listener can discriminate and evaluate. The
human ear remains far more sophisticated, especially when trained, than current computer
software dealing with timbre recognition or using computational auditory scene analysis
(Ellis 1986). And the issue of determining without error the perceptual salience of
patterns of peaks and dips in the spectrum for individual listeners beyond the level
achieved by the psychoacoustic models used in the development of codecs for sound
compression (Brandenburg 1993) remains unresolved. As a result, an accurate model of
the student in a timbral ear training course is beyond the scope of current research.
3.3 A monitoring system

The training method presented in this dissertation was motivated in part by the lack of information we have about the skills, knowledge, and listening strategies that are used during spectral adjustment tasks, and a desire to increase our knowledge in this domain. Such information can potentially be instrumental in developing better training tools. However, this is mostly non-verbal knowledge which listeners performing the task have difficulty to express verbally. Laske (1992) described a system that acquired such non-verbal knowledge about musical composition skills by simulating a typical compositional environment and implementing observation mechanisms in the software to gather data about the user's actions during the process of composing music. A similar approach was adopted for the present research.

Complex spectral adjustment tasks in which patterns of peaks and dips must be identified are seen as problem solving tasks (Anderson 1990). In listening tasks in which spectral modifications must be cancelled, the listener uses various problem-solving operators (e.g., parameter controls on a spectrum equalizer) to go from the initial state (the spectrally modified version of the sound sample) to the goal state (the version of the sound sample in which all spectral modifications introduced by the software have been removed).
Implementing this approach in a computer software allows these operators to become sensors that can be used to monitor and record the actions of the listener during the solving of an equalization problem. If the sensors can attach timestamps to these actions, then this additional data can be potentially useful for training purposes since the timing of interactions between the user and the software can be an important source of diagnostic information (Fox 1988).

3.4 Evaluation criteria

The criteria used to assess the skill level of listeners engaged in the evaluation of sound quality differ somewhat depending on the purpose and context of the evaluation. In listening tests for the evaluation of audio equipment, a sensible evaluation scheme will attempt to evaluate if a given listener can discriminate between stimuli in a consistent manner. However, as pointed out by Bech (1989), measuring the abilities of a listener to evaluate loudspeaker sound quality based solely on standard deviation measures on a set of judgements is not adequate. A listener with poor discrimination abilities might rate all loudspeakers under study equally and thus obtain a standard deviation of 0. The "loudspeaker test statistic $F_L$" (Bech 1990, 1991) appears to be a better and more accurate indicator that can be integrated into training software (Olive 2001). The $F$ test is calculated from data obtained from an ANOVA analysis of variance and measures the
Chapter 3: Training Methods and Issues

ratio of variance due to differences between the loudspeakers (plus random variation) over the variance due to random variation in a subject's responses (Bech 1989, p. 16). An F value higher than 1.00 thus indicates that the variance due to differences between loudspeakers is higher than the variance due to inconsistencies in the listener's evaluations.

Sound engineers must also be able to detect small spectral impairments in the frequency response of audio recordings as well as being consistent and reliable in their perceptual judgments. A measure of the listening skill level of audio engineers involves the additional idea of performance which in turn includes measures of accuracy and response time. Accuracy is a product of the ability of a listener to discriminate between two sounds since the final criterion the listener must use before registering an answer is whether the sound being spectrally modified is the same as or different than the reference. Response time is affected by:

- the robustness and extent of a listener's memory of timbre references
- the ability to direct or to discipline auditory attention to relevant areas of the spectrum, and
- effective listening strategies.
Chapter 3: Training Methods and Issues

The evaluation of a listener in the initial implementation of Timbral Ear Trainer relied only on the accuracy of the answers. For example, a listener would obtain a maximum score in a formant-matching task involving 3 resonances as long as the three center frequencies of the formants applied by the listener would match those applied by the software in the question.

Experience teaching timbral ear training has indicated repeatedly that in such formant-matching tasks involving octaves and 1/3-octave frequencies, most listeners will eventually reach the right answer, provided they are given enough time. However, one can reasonable conclude that a listener reaching the right answer in 40 seconds will display a higher level of performance than another listener reaching the same answer in 5 minutes. Response time thus seems to be a potentially useful indicator of a listener's performance. However, there can be more to a given response time than meets the eyes. A short response time can contain information about the robustness and reliability of the long-term memory for timbre references and can indicate the use of more useful strategies. The number of sound comparisons a listener performs during a given trial can also be a potentially very useful indicator of his ability to rehearse sound references in working memory. A related indicator of performance is the number of steps taken by the listener to reach the right answer. The better listener will take fewer steps and thus perform more quickly to arrive at the right answer. The difficulty of the spectral adjustment task can also be a factor in the assessment of a listener's ability. Two listeners
might perform equally well on an easy task but a more difficult task might reveal that one of the listeners performs better than the other.

### 3.5 Deriving a performance index

Considering the evaluation criteria described in the previous section, is it possible to derive a performance index that could be used as a global indicator of the level of expertise achieved by a listener on timbre evaluation tasks? Such an index is proposed in this thesis that takes into account both accuracy and response time.

Performance index is defined as a weighted ratio of matching/identification accuracy over response time:

\[
PI = \frac{S}{k(T_R)}
\]  

(1)

where:

- \( PI \) is the performance index,
- \( S \) is a measure of accuracy of the adjustment,
- \( T_R \) is the response time in seconds, and
- \( k \) is the weight applied to response time in the calculation of \( PI \)
Chapter 3: Training Methods and Issues

According to this equation, if accuracy (S) is zero, then performance is rated as null. Otherwise, overall performance will decrease as the response time denominator increases. $K$ can be varied between 0.1 and 1 depending on the weight assigned to response time.

3.6 How to cope with erroneous data

The monitoring of the actions of a listener during a listening task can produce meaningless, irrelevant data or "noise". Sources of noise should be minimized or otherwise handled to a certain degree to avoid as much as possible misleading indications that can potentially produce inappropriate responses from the software.

3.6.1 Listening fatigue

Listening fatigue will occur or manifest itself more or less rapidly depending on the initial physical and mental state of the listener, the difficulty of the exercises, the sound samples that are used and the frequency regions where spectral modifications are applied. It has been found that for broadband noise, auditory fatigue will be most pronounced between 4 and 6 kHz (Moore 1989). Listening fatigue can typically raise the threshold for the detection of timbre differences that would otherwise be not very hard to
Chapter 3: Training Methods and Issues

Fatigue can also lead to a lack of attention. Oftentimes, listeners themselves will detect they have reached a high level of fatigue after, for example, a series of consecutive mistakes on tasks that they usually can handle.

3.6.2 Lack of attention

My experience in teaching timbral ear training and performing spectral adjustment tasks suggest that attention is one of the most important factors affecting performance of a listener in the evaluation of timbre differences. As it has been discussed in Anderson (1990), attention is a limited cognitive resource. In a typical timbre evaluation task involving multiple spectral modifications, attention must be directed to specific spectral regions. Part of the training in timbral ear training is to develop disciplined auditory attention. The attention of a tired listener can be more easily grabbed by some irrelevant detail of the sound sample that will conceal more salient and relevant timbral features. Similarly to listening fatigue, a lack of attention can reveal itself by producing unusual mistakes such as confusing octave formants separated by 3 octaves, or missing an obvious timbre difference between the asked question and the user's answer. Lack of attention can also cause manipulation errors which are discussed below.
3.6.3 Human-computer interaction problems

The interface provided to the listener to interact with the software can potentially cause difficulties and introduce noise in the produced data. In absolute identification tasks in which the listener has a time limit to identify the center frequency of a formant, required mouse movements can add to the listener's actual response time and thus produce misleading data. In its current version, Timbral Ear Trainer uses the computer keypad mapped to the nine octave frequencies used in absolute identification tasks to let the listener enter answers. While this has significantly decreased interaction time overhead compared to the use of the mouse to click on a frequency value button, it limits, in the best cases, response time to approximately 0.5 second. Although such response times are not typically required in sound recording, loudspeaker design, or listening tests on sound quality, the ability to identify very quickly the center frequency of a spectral resonance implies that the listener needs very little sensory information to make the right assessment, which is in itself a measure of the robustness of the memory of these timbre references. I suspect that a more efficient data entry tool might allow the recording of shorter response times than the above. In timing more complex formant matching tasks where more manipulations are required from the user (e.g., adjusting multiple sliders), a greater amount of "interface usage time" is added to the actual response time the listener is capable of. This overhead can be minimized by providing user interface short cuts
whenever it is possible. The Keystroke-Level Model (Card et al. 1983) for quantitative analysis of interface designs has produced a set of timings for typical user interface gestures. For example, it has been found that it takes approximately from .12 to 1.2 seconds to tap a key, 1.1 second to point the mouse to a particular location on the screen, 0.4 seconds for a user's hand to move from the keyboard to a mouse or vice versa, and from 0.6 to 1.35 seconds to prepare mentally for the next action to perform. These numbers are derived from a simple model and vary in practice. But they nevertheless provide useful approximations.

3.6.4 Manipulation errors

The Timbral Ear Trainer software is not tolerant of errors in exercises involving the adjustment of equalization parameters. One reason for this is that the software cannot determine if a very small difference in say, gain or frequency settings, will result in a perceivable difference in timbre. One possible workaround is to develop means of preventing the user from making common manipulation errors. For example, a slider used to manipulate the center frequency of a resonance or dip can automatically be set to the closest octave or 1/3 octave value when its knob is released. The software can be designed to detect if the listener has forgotten to adjust the gain in at least one frequency band and either notify the user or simply disable the answer button until a gain (hence a
Chapter 3: Training Methods and Issues

timbre change) is applied to the sound. Similarly, most exercise parameters include frequency ranges within which values can be chosen. The software can be designed to detect out-of-range values and subsequently notify the user.

3.7 How to cope with uncertain perceptibility in contextual listening

This section discusses an issue that is difficult to address and solve within the scope of this dissertation. Timbral Ear Trainer does not implement any machine listening algorithm. Even if it did, there is no guarantee that a timbre matching judgment performed by a machine listener would not actually correspond to a timbre mismatch by a human listener. One fundamental requirement of software for timbral ear training is to avoid a situation where the software would tell the listener that two sound stimuli have identical timbres while in fact a difference in timbre could actually be heard.

Just noticeable differences can be measured in a laboratory under controlled conditions using controlled stimuli. Hearing thresholds will however vary depending on the audio reproduction environment (loudspeakers, room acoustics) and the audio material used during a training session. They will also be affected by the listener's attention level and physical/mental conditions. For example, the various results reported
Chapter 3: Training Methods and Issues

by Pollack (1963), Bücklein (1981), Green (1984), Moore (1989), and Farrar (1987), do not necessarily correspond to results that can be observed in the context of timbral ear training. Each sound stimulus used for training provides a different listening context that can affect the detectability of peaks and dips in the spectrum.

Similarly, neither the critical bandwidth formula proposed by Zwicker and Terhardt (1980), nor the ERB equation proposed by Moore and Glasberg (1983) can account for the variety of contexts that can be encountered in real-world listening. Trahiotis (1983) has pointed out that it is important to distinguish between "how well people typically do perform from how well they can perform in highly controlled and nearly optimal situations" (p. 63). In the Timbral Ear Trainer software, it is up to the instructor defining exercises to avoid including questions that may include spectral differences that are not perceivable.
Chapter 4: Implementation

This chapter describes the implementation of the rule-based monitoring system used for the training method introduced in section 3.3. An overview of the training program is first presented. Then the listening task used to illustrate the monitoring-based training method is described along with the interface tools that are provided for the user to interact with the software. The monitoring system architecture, its elements and their functions are described and discussed.

4.1 Training program overview

The training program is structured around weekly lessons presenting tasks of increasing complexity, each lesson building on the material covered in the preceding lessons. Each lesson comprises practice exercises and a self-test. There are two distinct and complementary parts in the training:

- one-to-one tutoring sessions with a human instructor where concepts are presented, tasks introduced, and difficulties, misconceptions, and questions addressed
- individual practice and training work with a computer
Chapter 4: Implementation

4.1.1 Development of memory for timbre

An important goal of the training is to develop memory for a set of octave resonance references centered from 63 to 16000 Hz. Center frequency values in Hz are initially used to identify the references. They are then complemented with the use of vowel labels between 250 Hz and 4 kHz. Once the octave references have been reasonably well mastered, 1/3-octave frequencies are introduced. A total of 25 1/3-octave frequencies are used from 63 Hz to 16 kHz. Identification of the third-octaves is exercised during the tutoring sessions. Working first on reduced frequency ranges, the students learn to recognize 1/3 octaves in relation to the octaves they have memorized. For example, 800 Hz can be recognized as being similar to 1 kHz (the vowel 'a') but a bit lower. 2.5 kHz is similar to 2 kHz (the vowel 'ε') but slightly higher. This method of partitioning the timbre categories allows the students to work around the limitation of 7 ± 2 categories we are usually able to manage in short-term memory (Miller 1956). They are encouraged to think first about the octave reference a given third-octave resembles to most and then determine if it is higher or lower. Some students become rapidly quite good at this task. A Q value of 2 is used for most of the training. The use of a medium Q helps each of the octave category and the vowels that are associated with some of them to have an identity distinct from the others. This is essential considering as discussed previously that the vowel-like resonances that are used are overlapping categories to a certain degree.
Chapter 4: Implementation

Higher Q values would yield a stronger sense of pitch rather than a sense of timbre and a lower Q value (e.g. 1 or below) would cause each category to lose too much of their identity. Gain values are initially set to +12 dB and are reduced progressively down to +3 dB for octaves and +6 dB for third-octaves.

4.1.2 Equalization adjustment tasks

The spectral adjustment tasks are simple at the beginning of the training (e.g. a single peak or dip at ±12 dB) but students become quickly able to evaluate patterns of 2 peaks and then patterns of peaks and dips. The maximum number of simultaneous peaks or dips at the end of the training is 3. In all these tasks, the students are encouraged to develop their ability to analyze the spectrum of the sound globally so they can identify all of the modifications. Uncertainty is introduced by constructing exercises in which the listener does not know in advance how many bands are modified for each question.

4.1.3 Perception of dips

The identification and recognition of dips in the spectrum are much more difficult to perform for human listeners than peaks (Bücklein 1981). As a result, the difficulty of
Chapter 4: Implementation

the exercises presented to the students is lower than what is asked for peaks. Students are trained to identify dips by analyzing the spectral content of the sound around a given dip. A dip in the spectrum will often reveal to the attentive listener the spectral region just above its center frequency. Students are also trained to "track down" dips by actively "looking for" the spectral region where an octave or 1/3 octave is missing.

4.1.4 Efficiency

In the second half of the training program, work on improving efficiency (being accurate and quick) is introduced by imposing a time limit for each question after which the user looses a percentage of the mark as specified in the exercise parameters. Work on speed is mostly done on absolute identification of octave formants. At the end of the training, response times of 1.5 seconds are achieved on single, 6-dB peaks but several students can reach response times between 0.6 and 1 second on many questions.

4.1.5 Exercise editor

An exercise editor is used to set the parameter specifications for each exercise, either for automatic generation of questions, or to specify a file containing pre-defined
Chapter 4: Implementation

questions. Exercise type, number of questions, number of tries allowed for each question, number of frequency bands to modify, range of values for center frequency, Q, and gain, whether to use octaves or 1/3 octaves, required response time, can be specified in the editor. Weights can be assigned to the frequency, Q, gain, response time and number of bands parameters. These weights are used during the evaluation of each answer. The exercise editor can be used to customize training programs for groups of listeners based on their needs. For example, the training content for student audio engineers and member of listener test panels can be quite different. Listening test subjects may be trained to evaluate broader peaks in the spectrum that are typically found in loudspeakers. The development of quick response time is less important for them than it is for sound engineers.

4.1.6 Sound Library

A CD-based sound library is used. Noise, speech in different languages, individual musical instruments, small ensembles and orchestral music is used. Musical sound samples are chosen from rock/pop, jazz, and classical music styles. For each sound sample, short sections are selected that present the best homogeneity both in terms of loudness and spectral content. The selected segment is looped during presentation using a
CD loop editor implemented in the software. These loop specifications can then be saved as sound samples in the library.

### 4.2 Listening task

Oftentimes in the work of sound recording, loudspeaker, and hearing aid engineers, there are spectral resonances or anti-resonances that need to be compensated for. In the audio recording chain, regions of the spectrum of musical instruments can be emphasized or cancelled through interactions with the room they are played in, and the microphones used to record them (Olive 1990). An important task of the sound engineer is to control this timbral shaping process. Some correction can be applied during the recording stage, others will be performed at the mixing stage. It is important that all components of a mix be well balanced spectrally to avoid masking problems that could degrade the transparence and intelligibility of the recording. The recording or mixing engineer who can detect and identify by listening where the unwanted resonances or anti-resonances are in the spectrum will be able to apply corrective measures quickly and effectively. Compensating for this type of distortion may require the recording engineer to apply some corrective equalization, or to change the microphones that are used, or else to modify the relative positions of sound source, microphones, and room boundaries.
During intermediary stages of loudspeaker design and development, a loudspeaker engineer can be faced with prototype versions displaying spectral irregularities that are detrimental to the overall sound quality of the product and thus will require corrections. Typically, listening tests will be conducted at these various stages to provide the engineer with subjective evaluations of the product being developed. The ability of the subjects participating in these listening tests and the loudspeaker engineer to detect and identify problematic spectral regions needing corrections through listening can facilitate greatly the development process by providing efficient and accurate assessments of the sound quality of the loudspeaker. The engineer may then have to change the design of the loudspeaker enclosure, or electronic components of the loudspeakers or revise some calculations. In both situations described above, being able to assess by ear where in the spectrum is the problem can be very helpful in finding a solution and re-evaluating the results efficiently.

The listening task used to illustrate the training method proposed in this dissertation is similar to the situations described above. In this task, the software applies spectral modifications to the sound sample. The task of the listener is to cancel this equalization by applying, for each spectral modification, the negative of the gain that was applied by the software at a given center frequency. For example, if a -6 dB dip is applied at 2 kHz, the listener must apply a +6 dB peak at 2 kHz. The final decision criterion
Chapter 4: Implementation

before registering an answer is whether the sound modified by the listener is the same as or different than the Flat version.

"Sound Sample" is defined here as the sound excerpt used for a particular exercise. There are three versions of the a given sound sample for each question:

Sound A: Contains the equalization applied by the computer. This is the question and its settings cannot be modified.

Sound B: In exercises involving equalization cancellation, sound B contains the original (audible but not visible) spectral modifications applied to the sound. This is the version the listener has to modify so that the spectral modifications are cancelled.

Sound Flat: This is the original, unprocessed version of the sound sample.

The listener is free to listen to and compare any version of the sound sample at any time and as many times as desired.
4.3 Audio tools

All sound processing is performed by a MIDI-controlled digital mixer, which provides a parametric equalizer for each input channel. The software tool used for the manipulation of the spectrum is thus a 4-band parametric equalizer (fig. 1). Each band covers the whole range of frequencies from 21 Hz to 18 kHz. The filters used to implement the equalizer are reciprocal so the result of applying a $-6 \text{ dB}$ dip on top of a $+6 \text{ dB}$ peak results in a perceptually flat spectrum. A second tool (the "Mixer") allows A/B/Flat comparisons as well as the adjustment of the listener level and balance (fig. 2). Finally, a simple audio CD interface allows the control of audio sample playback.

![Parametric Equalizer](image-url)

Figure 1. Equalizer controller window.
4.4 Monitoring architecture overview

This section presents an overview of the monitoring module's architecture (see fig. 3). The monitoring module is implemented using CLIPS (C Language Integrated Production System), a multi-paradigm programming language combining a rule-based language with object-oriented and procedural constructs (Giarratano & Riley 1994). The user interacts with the software through a collection of problem solving operators.
Chapter 4: Implementation

implemented in the user interface as controls, sliders, and buttons. The recording of these operators' usage over time contributes to the building of a context which describes the current state of the training procedure (what is the task being performed, what is the listener doing, etc.). This monitoring system offers two important advantages. First, it allows the training module to build a picture of the current training context without as noted previously building a dynamic cognitive model of the listener. Second, the observation mechanism can record potentially useful interaction data that can be used in future research to investigate the problem-solving strategies used by different listeners while solving similar tasks.

The Observer is the entry point of the module. It is the C++ object that provides the link between the user interface elements and the corpus of code written in CLIPS which implements the rule-based monitor. It collects information about the user's activity. The Context is the "working memory" of the module. It contains data about the current state of the TET world as captured by the Observer. The Monitor object attempts to derive from the raw events actions and strategies that are added to working memory. Monitor and Observer are complementary objects whose names may be somewhat confusing. The Observer (written in C++) sends raw events to the CLIPS engine where they are stored in working memory (the "current context"). The Monitor (written in CLIPS) converts (at least tries to) these raw user steps into actions and possibly strategies. Both Observer and Monitor contribute to the building and updating of the working
memory (context). The knowledge base (KB) contains heuristics, procedural and declarative knowledge about training equalization removing tasks and spectrum equalization. The knowledge base uses the inference engine provided in the CLIPS package to return feedback to the user through user interface objects.

Figure 3. Observer architecture. The Observer object collects data from the user interface operators and adds the information to the software's working memory. Using data contained in the knowledge base, the Monitor object attempts to "understand" what the user is doing in the context of the current spectral adjustment task by deriving actions and strategies from the raw user events with the help of the inference engine. The output of the monitor object updates the current context and when appropriate can provide information to the user. The right-hand side block is the internal observer engine; the left-hand side is its interface to the outside.
4.5 Observing listeners

During a timbral ear training tutoring session, the instructor can ask at various points why the listener performed a particular step, which part(s) or aspect(s) of the sound were used to arrive at a particular answer, where was the auditory attention directed to. In turn, the student can comment on difficulties that are encountered, asking about strategies to overcome them, wondering if the approach that was taken in a specific situation is a good one. These dialogs between the instructor and the student are invaluable and provide a wealth of information about the way a particular question was approached, etc. It helps the student improve the strategies used during listening, it provides the instructor with a better understanding of the learning/training habits of the student.

Between tutoring sessions, the listener works alone on exercises with the ear training software. This section describes how the mechanism to observe the listener that was presented in the previous section works.

4.5.1 Tracing a user's actions and strategies

The main function of the Observer object is to gather information about what a listener is doing in the course of solving an EQ Removing task. The Observer installs
Chapter 4: Implementation

"sensors" at all interaction points of the software. When an exercise starts, the sensors are activated, together with a microsecond timer. The Observer records all user interactions with the software as basic raw events. There are various types of raw events, each type of event having its set of attributes. All event types have a name attribute such as "button-step" or "eq-fader-step". Event types are defined in various categories. All raw events have a "time-stamp" attribute. Most have a "value" attribute. Other events have additional attributes such as "band" and "parameter" for the "eq-fader-step" events which identify which frequency band was modified and which of the center frequency, Q, or gain parameter was manipulated.

The list of defined raw events includes:

- **button-step**: A generic button step event
- **eq-button-step**: A specialized version of the above with additional attributes related to EQ
- **menu-step**: Any menu item selection
Chapter 4: Implementation

eq-fader-step:
Each of center frequency, Q, and gain parameters in the equalizer are controlled using a fader.

mixer-fader-step:
This type includes changes of the listening level or the level balance between samples A, B, and Flat.

eq-vowel-step:
Event generated by the use of a vowel operator which adjusts the center frequency of a band to the corresponding vowel.

peak-dip-step:
Event generated by an operator allowing the user to specify if a timbre modification in one of the bands is perceived as a peak or a dip in the spectrum.

An important function of the Monitor module is to attempt to group these individual events into meaningful actions. Rules are defined to add the following actions to the current context:

listen-to:
Any time a sound sample is selected for audition, this action is added to the actions history and context.
Chapter 4: Implementation

frequency-adjustment

Q-adjustment:

gain-adjustment:

Whenever a parameter in one of the frequency bands of the equalizer is changed, the appropriate action is generated.

vowel-identification:

Each of the frequency band in the equalizer tool provides a set of vowel buttons that the user can choose instead of adjusting the frequency fader for the corresponding frequency value in Hz.

listening-level-adjustment:

Preferred listening level may vary from one user to another or depending on the sound sample being played. Some listeners may be tempted to increase the listening level when the resonance magnitude in an exercise is low.

AB-balance-adjustment:

The level of the sound samples (A, B, and flat) can be adjusted separately.

Finally, the Monitor attempts to derive (or infer) listening strategies from the pattern matching of the current context and recent user actions. Defined listening strategies include:
Chapter 4: Implementation

sound-comparison:

A sound comparison may be used by the listener to determine where in the
spectrum a peak or dip is applied. It is also used to determine if the two sound
samples are the same or different.

eq-adjustment-with-aural-feedback:

eq-adjustment-without-aural-feedback:

Spectrum adjustments can be made either while listening in real time to the timbre
changes they produce (when the listener is listening to sound B) or without
listening to the timbre changes as they occur.

vowel-identification:

Within the relevant frequency range, some listeners will, sometimes
spontaneously, hear the vowel-like sound associated with a given resonance's
center frequency.

rapid-comparisons:

slow-comparisons:

The time a listener takes to listen to each version of a sound sample when
comparing two versions can have a strong impact on the ability to detect a
difference (if there is one) between the two. Rules are implemented that allow the
software to detect if two sounds may be compared too quickly to allow a reliable
detection of any difference. However, the comparison speed threshold at which it
becomes difficult to detect a timbre difference may depend on the listener attention, listening skills, and programme material.

**search-resonance-by-sweeping:**

A usually effective strategy in which the center frequency of a resonance is found by sweeping the center frequency of an anti-resonance across the frequency range until the original resonance is removed. This requires that the listener is able to remember and rehearse "internally" the target resonance while the center frequency of the anti-resonance is swept across the allowed frequency range.

### 4.5.2 Problem solving operators

When the listener works alone at the computer, the internal thoughts underlying the problem solving steps he performs to achieve a given answer are lost unless either the user interface provides elements allowing the listener to externalize otherwise covert thoughts or else the knowledge base is sophisticated enough to be able to determine from the current context what the listener is doing. The approach presented in this dissertation provides the CLIPS inference engine with as much direct information as possible through the use of additional controls that allow the listener to specify some of the internal thoughts as they arise. These controls in turn generate events that the Observer captures
and adds to the current context. Changes in the context may trigger rules that will use the new data to infer additional information about what the listener is doing.

Every controller in the user interface is a potential problem solving operator. In the present work, a problem solving operator is a user interface element the listener can manipulate to change the state of the TET world towards the solution to a problem. In this sense, a frequency fader is a problem solving operator as it allows the user to directly modify the spectrum of the sound sample. The buttons that switch between the different versions of the sound sample are also problem solving operators as they allow the listener to compare the question with his answer to verify his progress or the validity of his adjustments.

Other operators allow the listener to decompose a given task into smaller and simpler sub-tasks. Let's take for example the EQ removing task described in section 4.2 above. For each of the three spectral modifications applied to the sound, the listener must decide if it is a peak or a dip in the spectrum. The listener can click, under each gain fader, on the radio buttons "P" or "D" to indicate a first impression or assessment. The listener can also think either in terms of vowels (if the range applies) or in terms of frequency value in Hz. A set of vowel buttons is available to the user. The choice is applied to the frequency fader and recorded by the Observer.
Chapter 4: Implementation

4.5.3 Event trace history

A trace of the individual user steps generated during a work session is stored in traces files. A trace file represents the script of a session displaying sequentially the steps and actions performed by the listener. For each question, the data describing the applied equalization is provided followed by a description of all user steps. Each step has a timestamp so the timing of interactions is remembered. The trace file is in a human readable format so the unfolding of a user's actions during problem solving can easily be followed. Which frequencies were tried, when and how many sound comparisons were performed, and the pace of the actions. For example, a listening strategy consisting in seemingly random adjustments can be seen in the following excerpt of a trace file where the listener produces a long sequence of resonance adjustments using the vowel tool within a short period of time. It appears that the listener is trying all possible vowel answers without knowing where in the spectrum the right one could be. The listener systematically tries all vowels repeatedly in the same band then goes on to adjust another band. Each set of 5 vowels are tried first as peaks, then as dips in the spectrum with comparisons in between adjustments. The numbering of each line corresponds to the numbering of steps for the whole session, including practice.

1009. ...
1010. Button "Mixer Sound B" used at 19:04:17
1011. Button "Mixer Sound Flat" used at 19:04:18
1012. Button "Mixer Sound B" used at 19:04:20
1013. Button "Mixer Sound Flat" used at 19:04:20
1014. Button "Mixer Sound B" used at 19:04:21
1015. Button "Vowel i" in band 3 used at 19:04:27
1016. Button "It's a peak" in band 3 used at 19:04:28
1017. Button "Mixer Sound Flat" used at 19:04:30
1018. Button "Mixer Sound B" used at 19:04:30
1019. Button "Vowel e" in band 3 used at 19:04:32
1020. Button "Mixer Sound Flat" used at 19:04:34
1021. Button "Mixer Sound B" used at 19:04:35
1022. Button "Vowel a" in band 3 used at 19:04:36
1023. Button "Mixer Sound Flat" used at 19:04:36
1024. Button "Mixer Sound B" used at 19:04:37
1025. Button "Mixer Sound Flat" used at 19:04:38
1026. Button "Vowel o" in band 3 used at 19:04:39
1027. Button "Mixer Sound B" used at 19:04:40
1028. Button "Mixer Sound Flat" used at 19:04:41
1029. Button "Vowel u" in band 3 used at 19:04:42
1030. Button "Mixer Sound B" used at 19:04:42
1031. Button "Mixer Sound Flat" used at 19:04:44
1032. Button "It's a dip" in band 3 used at 19:04:46
1033. Button "Mixer Sound B" used at 19:04:46
1034. Button "Mixer Sound Flat" used at 19:04:48
1035. Button "Mixer Sound B" used at 19:04:49
1036. Button "Vowel e" in band 3 used at 19:04:49
1037. Button "Mixer Sound Flat" used at 19:04:50
1038. Button "Mixer Sound B" used at 19:04:51
1039. Button "Vowel a" in band 3 used at 19:04:53
1040. Button "Mixer Sound Flat" used at 19:04:53
1041. Button "Mixer Sound B" used at 19:04:54
1042. Button "Vowel e" in band 3 used at 19:04:55
1043. Button "Mixer Sound Flat" used at 19:04:55
1044. Button "Mixer Sound B" used at 19:04:56
1045. Button "Mixer Sound Flat" used at 19:04:57
1046. Button "Vowel i" in band 3 used at 19:04:58
1047. Button "Mixer Sound B" used at 19:04:59
1048. Button "Mixer Sound Flat" used at 19:05:00
1049. Button "Vowel a" in band 3 used at 19:05:01
1050. Button "Mixer Sound B" used at 19:05:01
1051. Button "Mixer Sound Flat" used at 19:05:03
1052. Button "Mixer Sound B" used at 19:05:04
1053. Button "Mixer Sound Flat" used at 19:05:05
1054. Button "It's no gain" in band 3 used at 19:05:06
1055. Button "Mixer Sound B" used at 19:05:07
1056. Button "Mixer Sound Flat" used at 19:05:08
1057. Button "Mixer Sound B" used at 19:05:15
1058. Button "It's a dip" in band 3 used at 19:05:19
Another example of an inefficient strategy is illustrated below by a long sequence of comparisons between B and Flat in which the listener seems to have difficulty determining if the two sounds are the same or different. Notice that the timestamps indicate very rapid comparisons which give little time to the hearing system to register (integrate) the timbre of each sample:
Finally, below is an example of efficient problem solving for a question in which there were peaks at 1 and 16 kHz. The listener starts by comparing carefully B and Flat, then quickly adjust both frequencies and carefully compare the final answer with Flat before clicking on OK, 47 seconds later.

4.5.4 A listener’s profile

The software builds and maintain a record file for each listener. The profile contains general information about the listener and data used in the management of the training. For example, the profile allows the listener to know what lessons and exercises
have been completed and what remains to be done. The state of an interrupted exercise is saved in the profile so that it can be restored and work resumed the next time. The profile also contains a list of results for each lesson the user worked on and cumulative data about the use of the software during exercises. The data include, for each answer:

- Data describing the question
- Data describing the subject's answer
- Raw score
- Number of steps taken by the user before reaching the answer.
- Response time
- Number of sound comparisons.
- Listening level
- Number of vowel identifications
- Number of times the "flat" reference was selected
- Number of times the A and B versions of the sound were selected
- Number of times the listener performed a reset of the equalization settings
- Number of times the bypass buttons were used
- Number of frequency adjustments
- Number of gain adjustments
Chapter 4: Implementation

For a given exercise, a performance index, as described above, and the total number of questions on which the data is calculated is computed. As will be discussed in the next chapter, these data can provide informative metrics on the performance and skill level of listeners.
Chapter 5: Validation

5.1 Introduction

The listening skills related to the identification and evaluation of timbre are believed to be developed over an extended period of time by audio engineers, without going through formal aural training such as described in this dissertation (Letowski 1985). The listening experiment reported in this chapter was motivated by the following question: How would the performance of student audio engineers with little real-world experience but formal timbral ear training (the experimental group) compare, on an advanced equalization task, with the performance of experienced audio engineers with no such formal training (the control group)? To my knowledge, there has not been any prior published work on such comparisons.

Two corollary questions were then asked: Did the tasks on which the subjects in the experimental group were trained involve listening skills that are used by professional audio engineers in the course of their work? Which parameters would reveal the greatest number of differences and similarities between the two groups?

The general hypothesis in response to these questions was twofold. First, it was hypothesized that the listening task presented to the subjects did involve listening skills
that professional audio engineers, in the absence of specific training, develop throughout their career. If the training had developed very specialized skills not relevant to work in the audio field, then 1) the test might have been too difficult for the control group and 2) the relevance of the training would have been subjected to serious questioning. It was thus expected that listeners in both groups would obtain good scores on the task. Second, it was hypothesized that overall performance would be superior in the experimental group than in the control group, even though subjects in the latter had significantly more professional listening experience as well as significantly more experience making equalization adjustments.

5.2 Test subjects

10 subjects participated in the listening test, with 5 subjects in the experimental group and 5 in the control group. All subjects were paid for their participation. The 5 subjects in the experimental group ("student subjects") were all enrolled in the Technical Ear Training course at McGill University, as part of the program for their Master’s degree in Sound Recording. Subjects in this group reported on average 2.4 years of training in audio and 4.5 years of experience working in the audio field. All of them had similar current audio activities through their common curriculum of study at McGill. The student group was therefore, at least on these aspects, fairly homogeneous.
Chapter 5: Validation

The 5 subjects in the control group were practicing audio professionals without prior formal aural training. Subjects within this group had on average 19 years of experience working in the audio engineering field and 2.2 years of training in audio. Their current professional activities in audio were more heterogeneous and specialized than in the experimental group with work areas including producing, recording musical performances, mixing, editing, and post-production of feature film sound tracks.

<table>
<thead>
<tr>
<th>Experience</th>
<th>N</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Standard Error</th>
<th>95% Confidence Interval for Mean</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>5</td>
<td>19.00</td>
<td>4.4159</td>
<td>1.9748</td>
<td>13.5170</td>
<td>24.4830</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>5</td>
<td>4.50</td>
<td>2.0616</td>
<td>0.9220</td>
<td>1.9402</td>
<td>7.0598</td>
</tr>
<tr>
<td>Total</td>
<td>10</td>
<td>11.75</td>
<td>8.3041</td>
<td>2.6260</td>
<td>1.9402</td>
<td>5.8096</td>
<td>17.6904</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>AUDIO TRAINING</th>
<th>N</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Standard Error</th>
<th>95% Confidence Interval for Mean</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio Experience</td>
<td>0</td>
<td>5</td>
<td>2.20</td>
<td>1.6432</td>
<td>0.7348</td>
<td>0.1597</td>
<td>4.2403</td>
</tr>
<tr>
<td>Audio Training</td>
<td>1</td>
<td>5</td>
<td>2.40</td>
<td>0.5477</td>
<td>0.2449</td>
<td>1.7199</td>
<td>3.0801</td>
</tr>
<tr>
<td>Total</td>
<td>10</td>
<td>2.30</td>
<td>1.1595</td>
<td>0.3667</td>
<td>1.4705</td>
<td>3.1295</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Table 1. Means, std. deviations and errors, and confidence intervals for audio experience and audio training. The "0" and "1" in the second column from the left correspond to the control and experimental (students) groups respectively.

Table 1 displays the means (in number of years), standard deviations, standard errors and confidence intervals for the audio experience and audio training characteristics of the two groups. The data reveal on the one hand that the control subjects had more audio experience than the trained subjects but that the number of years of academic audio training was similar. An ANOVA of these two parameters (table 2) indicates that the
difference in the number of years of experience in audio between the two groups is very significant \((p \leq 0.001)\) while the amount of audio training is not. In light of the research questions that were stated at the beginning of this chapter, the dissimilarity regarding audio experience allows a more rigorous test of the training method presented in this thesis since subjects in the control group have the advantage of experience. The effect of a short period of structured, theory-based training is thus measured against the effect of many years of hands-on experience.

<table>
<thead>
<tr>
<th>EXPERIENCE</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>525.625</td>
<td>1</td>
<td>525.625</td>
<td>44.263</td>
<td>0.000</td>
</tr>
<tr>
<td>Within Groups</td>
<td>95</td>
<td>8</td>
<td>11.875</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>620.625</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUDIO TRAINING</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between Groups</td>
<td>1.00E-01</td>
<td>1</td>
<td>1.00E-01</td>
<td>0.067</td>
<td>0.8</td>
</tr>
<tr>
<td>Within Groups</td>
<td>12</td>
<td>8</td>
<td>1.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>12.1</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. ANOVA of audio experience and audio training for the two groups.

5.3 Stimulus

The programme material used for the test was "Gaslighting Abbie" from the popular music group Steely Dan (Dan 2000). The pop song was selected for the high quality of its recording (winner of the 2000 Grammy Award for "Best Engineered
Chapter 5: Validation

Album/Non Classical), and has a wide and fairly well balanced spectrum allowing the elements of the mix to blend well while retaining a high level of spaciousness and sonic separation. None of the subjects were familiar with the selection. A 49-second segment of the sound sample (from 0:40:05 to 01:29:50 in min:sec:frames units) was selected for the best possible homogeneity and rich spectral content. The stimulus was presented in the small listening room used for the Technical Ear Training course, using two Genelec 1030 loudspeakers arranged in a standard equilateral triangle stereo listening configuration.

5.4 Procedure

5.4.1 Task

The test simulated a task common in sound recording practice: compensating for spectral irregularities by canceling peaks and dips in the spectrum using a parametric equalizer, as described in section 4.2 above.

Table 5-3 shows the equalization data for the 15 trials in the test. In each trial, from 1 to 3 spectral modifications were applied to a sound sample. The number of modifications presented on each trial was unknown to the subjects. The 0 values in the table indicate frequency bands that were not modified. Each modification was either a +6
dB peak or a -6 dB dip. In trials in which multiple modifications were applied, any combination of peaks and dips was possible. A pilot experiment was conducted to determine optimum Q value, frequency resolution and range to use in the test. A medium Q value of 2.0 (not included in the table) was chosen and used in all trials, and was known to the subjects. Frequency values were chosen among octaves between 125 Hz and 16 kHz. The 63 Hz value was not included because it was assessed that there were, for the stimulus being used, some interactions between 63 and 125 Hz (e.g., applying a peak to one and a dip to the other) that could be too hard to hear to be included in the test. The use of octaves instead of 1/3 octaves was also motivated by the need to set the difficulty of the test at an appropriate level (not too easy nor too hard, but challenging for both groups.

Subjects had to determine, for each trial:

a) the number of spectral modifications that were applied to the sound,

b) whether each modification was a peak or a dip, and

c) the center frequency value of the modification.

For each of the spectral modifications in a given trial, the task was to apply the negative of the gain applied in the original equalization, with identical center frequency. The Q, being fixed, was set automatically by the software.
Table 3. Trial List. The Q value is not included in the table since it is always equal to 2.0. Bands in which the gain and frequency values are 0 indicate bands that were not modified in the corresponding trial.

The test was, in essence, a matching task in which the subjects had to modify the comparison stimulus (each trial equalization) until it matched the "flat", unprocessed reference, which was the same for each trial.

Subjects could compare the sound they were modifying with the flat reference as many times as they needed to. For each trial, the perceptual criterion for a right answer was that the flat and the user-modified version of the sound sample had the same timbre. Spectral modifications were performed using the parametric equalizer described in section 4.3.
Subjects were informed that they would be evaluated both on the accuracy of their answers and their response time. They were instructed that they should aim at being accurate, as quickly as they could (see Appendix 1 for the instructions to the subjects). For the purpose of the test, it was important that all subjects completed all trials. Therefore, no time limit was imposed on the completion of the test.

The presentation order of the trials was randomized for each listener to cancel out any contextual or order effects that might have been introduced by a specific sequencing of the trials. Two functions were implemented in the software to further minimize the possibility that subjects might evaluate a given trial based in part on their memory of the previous one. First, every trial started with the presentation of the reference sound. Therefore, all equalizations were first evaluated in comparison to the unprocessed sound. Second, the sound was muted between trials. This had the additional benefits of preventing subjects from practicing between trials and allowed them to rest if they needed to.

All subjects were given a practice period prior to starting the test to familiarize themselves with the operation of the software, the test procedure, and the stimulus. There were a total of 6 practice trials each subject could go through. Subjects proceeded to the test when they felt ready to start. The student group was also offered this practice period although they were already familiar with the software and the task. However, they were not familiar with the test procedure nor with the stimulus.
Chapter 5: Validation

Subjects were not given any feedback about their performance during the test other than the aural feedback inherent in the comparison of the equalized and flat versions of the stimulus.

5.4.2 Score evaluation method

Each trial in the test was assigned a raw score between 0 and 1. All raw scores were then multiplied by 100 to yield a percentage raw score from which further analysis was performed as will be described shortly. Grading of individual trials obeyed the following rules:

For each trial, weights of 65% and 35% were assigned to the frequency and gain parameters respectively. Response time was not used in the calculation of raw scores but was used later in the computation of the performance index. On each trial the two weights were divided up based on the number of bands to adjust. For example, in a trial with a single modification, frequency was worth the full .65 mark. In trials with 2 frequency bands to adjust, each right frequency was worth 0.325. The same calculation was applied to the gain value. Table 4 lists the parameter weights based on the number of modified bands.
Table 4. Evaluation weights. For each of the two parameters, the weight is divided by the number of equalized bands to adjust.

This method took care of trials for which subjects had not modified enough bands. For trials in which listeners had modified more bands than there were in the question, a fixed value of 0.2 was subtracted from the mark for each extra bands.

5.5 Results and discussion

The performance of the two groups was compared along several parameters, from global score to the usage of specific problem solving operators. The results of the data analysis for each parameter are presented and discussed separately. The findings are summarized at the end of the chapter. The main results are listed in tables 5 and 6 on the next two pages.
Table 5. Means, standard deviations and errors, and confidence intervals for the performance parameters that were analyzed. The "0" and "1" in the second column from the left represent the control and experimental groups respectively, each group containing 5 subjects.
Table 6. ANOVA of the performance parameters under study.
5.5.1 Score

The mean scores obtained by the control group and the experimental group were 78.8% correct (s.d.=16.8) and 96% correct (s.d.=2.98) respectively (fig. 4). A one-way analysis of variance revealed that the score difference between the two groups was statistically significant ($F(1,9) = 5.08, p \leq 0.05$).

![Figure 4. Average score (%) and standard deviations for the two groups.](image)

The histogram of the mean scores for the 10 subjects (fig. 5) shows a negatively skewed distribution, indicating as was expected that most subjects obtained a high score.
Figure 5. Distribution of mean scores (%) for all subjects. The subject with a score of 100% is in the interval 100-110.

However, the distribution of scores plotted separately for the two groups (fig. 6 and 7) reveals that while the scores of the student group still show a negatively skewed distribution with a high mode, the control group's scores have a wider distribution. These two histograms show that the control group is more heterogeneous than the student group with the lowest score within the 50-60 interval and the highest within the 90-100 interval while the student subjects all have scores above 90%. This may be related to the
homogeneity effect of the training observed by Quesnel (1991). It may also illustrate the asymptotic effect of training on skill level.

Figure 6. Distribution of scores for the experimental group. All scores are within the 90-100% interval, hence the reduced width of each bar compared to the previous graph.
Figure 7. Distribution of scores for the 5 control subjects. A majority of scores are above 80 but two subjects obtained lower scores.

One additional analysis of the score data was performed. I counted, for each subject, the number of 100% correct answers and then computed the ratio of these to the total number of trials (15). Students had significantly ($F(1,9) = 6.58; p \leq 0.03$) more individual (on a trial by trial basis) perfect scores than the subjects in the control group. The results may indicate that the students aimed more at perfect scores than the controls. It may also indicate that students were better at determining when their answers were not right. These results are listed in tables 7 and 8.
Chapter 5: Validation

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Standard Error</th>
<th>95% Confidence Interval for Mean</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lower Bound</td>
<td>Upper Bound</td>
<td></td>
</tr>
<tr>
<td>Controls</td>
<td>5</td>
<td>8.00</td>
<td>4.1833</td>
<td>1.8708</td>
<td>2.8057</td>
<td>13.1943</td>
<td>4.00</td>
</tr>
<tr>
<td>Students</td>
<td>5</td>
<td>13.00</td>
<td>1.2247</td>
<td>0.5477</td>
<td>11.4793</td>
<td>14.5207</td>
<td>12.00</td>
</tr>
<tr>
<td>Total</td>
<td>10</td>
<td>10.5</td>
<td>3.9229</td>
<td>1.2405</td>
<td>7.6937</td>
<td>13.3063</td>
<td>4.00</td>
</tr>
</tbody>
</table>

Table 7. Means, standard deviation, standard error and confidence intervals for the number of perfect scores obtained by the subjects in the two groups.

<table>
<thead>
<tr>
<th></th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>62.5</td>
<td>1</td>
<td>62.5</td>
<td>6.579</td>
<td>0.03</td>
</tr>
<tr>
<td>Within Groups</td>
<td>76</td>
<td>8</td>
<td>9.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>138.5</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 8. ANOVA of the mean number of perfect scores for the two groups.

5.5.2 Response time

It was initially hypothesized that response time means would be lower for the student group than for the control group. However, the difference between the two groups was found to be small (0.54 minutes) and wasn't statistically significant. Figure 8 shows the bar graph with standard deviations for response time.

A quick response time is desirable only if the adjustments are correct. It may be that response time by itself is not a very good indicator of performance since a subject can
give wrong answers very quickly. It was not meaningful to look at response time values only for those trials that had received perfect scores. Many subjects obtained partial scores on several trials and it was not practical in these cases to separate response time values for the correct and incorrect parts of a given trial.

5.5.3 Listening and usage times

Listening time, for each subject, is the sum of the response times for the 15 trials.

Usage time includes both listening time and time elapsed in between trials over the
duration of the test. Means of these sums were computed for the two groups. Control subjects took on average 71.6 minutes to complete the test while student subjects took 40.4 minutes. The difference between the two groups did not meet the $p \leq .05$ significance level criterion but nevertheless indicated a strong trend with a $p$ value of .09. The mean listening time was 47.6 (s.d. = 15.0) minutes for the control group and 34.8 (s.d. = 9.1) minutes for the experimental group. This is in agreement with the small differences in response time means reported above. The difference between usage time and listening time corresponds to time periods in between trials during which subjects could rest. A common observation expressed by subjects in the control group was that the task was much more tiring than they expected and it can be hypothesized that they needed more rest periods than the trained listeners. Figures 9 and 10 show bar graphs for listening and usage times.

5.5.4 Number of steps

The mean number of step values include all sound adjustments (frequency adjustment, vowel identifications, gain adjustments, listening level, etc). It was hypothesized that the student subjects would not perform as many sound adjustments as the control subjects because presumably, they would be better at analyzing the spectral differences and would apply the right adjustments more directly than the control group.
who would use more trial-and-error strategies and successive approximations to achieve a correct answer. The mean obtained for the control group was 89 (s.d. = 22.4) steps while the mean for the student group was 58 (s.d. = 17.0). The difference in the two groups was found to be significant (F(1,9) = 6.08, p ≤ .05). Figure 11 shows a bar graph with standard deviations of the mean number of steps for the two groups.

Figure 9. Mean listening time for control and student subjects.
5.5.5 Sound comparisons

The number of sound comparisons a subject makes during a trial can reveal, at least in part, the extent to which the subject relies on memory to perform sound comparisons. It was thus hypothesized that the trained subjects would compare B and Flat less often on average than the untrained subjects. The results revealed a trend in that direction ($p = .09$) with means of 64.2 and 40.9 for the control and student group.
respectively. The value obtained for the trained subjects was higher than anticipated. A potential explanation might be that during the 6-month training period, subjects in the

![Figure 11. Mean number of steps for the control and student subjects.](image)

experimental group received continuous feedback as they worked through the exercises. They may have come to rely too much on this feedback. The listening test described here was the first time they did not receive such feedback and reinforcement. It may have had an impact on their overall performance, and on the number of sound comparisons they performed. The lack of feedback increases the level of uncertainty and most trained
subjects may have tried to compensate by performing more sound comparisons than usual. Figure 12 shows a bar graph of group means for the sound comparisons.

![Bar graph showing mean number of sound comparisons for each group of subjects.]

Figure 12. Mean number of sound comparisons for each group of subjects.

5.5.6 Vowels

The usage of the vowel buttons to specify a frequency was examined. The mean values obtained for the two groups were similar and the difference between the two groups was not significant. Subjects were instructed to use these buttons only if they were familiar with the relationship between vowels and resonance frequencies. All control
subjects said they were not familiar with the concept but most of them nevertheless did use them. In informal comments after the test, some control subjects indicated they had used the vowel buttons as shortcuts to try different center frequencies. The user interface could have been designed differently to address this issue.

5.5.7 Gain and frequency adjustments

The mean number of gain adjustments include both the use of the gain faders and the peak/dip buttons. The student subjects made slightly less gain adjustments than the control subjects but the difference was not significant. The mean number of frequency adjustments included the use of the frequency faders and the vowel buttons. The difference between the two groups was significant \( p \leq 0.05 \) with means of 12.80 and 8.98 for the controls and students respectively and standard deviation values of 2.65 and 1.03. There were only two gain values to choose from compared to 8 frequency values. Since there was a single gain value for each of the peak and dip categories, the decision about the gain adjustments was based more on deciding if a given modification was a peak or a dip. Subjects among the control group did confuse peaks and dips on occasion. None of the student subjects confused the two. The number of frequency adjustments reveals better the effect of training (fig. 13).
Figure 13. Mean number of frequency adjustments (combining the use of frequency faders and vowel controls) with standard deviations.

5.5.8 Listening level

The mean listening level was examined for the two groups as a potential indicator of the use of level to emphasize the perceptibility of the lowest (125 Hz) and highest (16 kHz) frequencies. The levels indicated below are level settings on the output bus fader of the mixer and not measured with a sound level meter. They are therefore used here as comparative values between the two groups and not as absolute values. The control subjects listened on average at a sound level 3.5 dB higher than the student subjects but
the difference between the two groups was found to be not significant \((p = .17)\). Please refer to figure 14 for a bar graph for the group means for listening level.

![Bar graph showing mean listening levels for controls and students](image)

**Figure 14.** Mean listening levels for the two groups. The dB values correspond to the output fader value on the mixing console.

### 5.5.9 Performance index

Finally, the performance index (as described in section 3.5) of the two groups was computed and compared. The control group obtained a mean PI value of 55.8 while the student group obtained 87.7, with values ranging from 41.4 to 123. The difference
between the two groups was found to be significant at the $p \leq .05$ level. The standard deviation value for the student group was higher (23.2) than the value obtained by the control group (13.5). This can be explained by the fact that the performance index equation used in this research does not take into account the difficulty of the equalization settings. A better listener will presumably be quicker at answering easier questions than harder ones and hence will obtain for these questions a higher PI value. Figure 15 shows a graph of the group means for performance index.

Figure 15. Mean performance index values for the control and experimental groups.
5.6 Summary

The results from the listening test described in this chapter indicate that overall, the student subjects performed significantly better than the control subjects as assessed by a number of different measures. Out of the 11 performance indicators under study, 4 were significant at the $p \leq 0.05$ level (score, number of steps, frequency adjustments and performance index). In addition, the analysis of the number of sound comparisons revealed a strong trend ($p = 0.9$) in the direction of higher performance of the student subjects.

The evaluation method of the subjects' answers allowed scores ($s$) to be incorrect ($s = 0$), partially correct ($0 > s < 1$) or correct ($s = 1$). Student subjects were significantly better ($p \leq 0.05$) at getting perfect scores than the control subjects. The variance in scores was much smaller for the student subjects than for the controls, in addition to their mean scores being higher. The findings reported here also indicate that response time taken alone is not a very good indicator of performance but instead must be examined in combination with accuracy, which the performance index measure developed in this research does.

The control group subjects had substantially more professional listening experience than the student subjects and this difference was highly significant ($p \leq 0.001$). Yet, the students outperformed them. In addition, the subjects in the two groups
had received a similar number of years of audio training. This strongly suggests that the technical ear training course the student subjects took as part of their audio training contributed substantially to their higher performance level.
Chapter 6: Conclusions

6.1 Summary

6.1.1 Training program

A new training method for the development of timbre memory and timbre evaluation skills has been developed. The method is implemented as a combination of individual computer-assisted work and one-to-one tutoring sessions with an instructor. Timbre modifications are produced by generating patterns of peaks and dips in the spectrum of sound stimuli. Each peak or dip is characterized by its center frequency, Q, and magnitude.

Listeners learn to develop their long-term memory for a set of nine timbre references based on octave resonances between 63 Hz and 16 kHz. They learn how to associate vowel qualities to the resonances within the vowel range (250 Hz to 4 kHz) and how to evaluate 1/3-octave resonances in relation with the octave set. Listeners also learn to discriminate between peaks and dips and work on recognizing center frequencies of dips using various strategies depending on the location of the dips in the spectrum. Listeners develop listening skills to recognize patterns of multiple peaks and dips applied
Chapter 6: Conclusions

to sound stimuli. Up to three spectral modifications are applied at a time and perceptual acuity is refined during the course of the training, with magnitudes ranging from +12 dB at the beginning of the training down to 3 dB at the end of the six-month period.

Accuracy of the timbre adjustments is emphasized throughout the training and efficiency is developed as well in the second half of the training, once listeners have acquired a stable set of timbre references. Efficiency is defined as being accurate as quickly as possible. Response times below 1 second are achieved at the end of the training in the identification of octave peaks.

6.1.2 Measurement of performance level

An important objective of the work presented in this dissertation was to examine how the expertise of a listener in the task of evaluating timbre can be assessed and measured. The performance of a group of 5 subjects who had taken the timbral ear training course at McGill was compared with a group of 5 subjects who had not taken the training but had considerably more professional listening and spectral equalization experience than the trainees. Various performance indicators were examined. Score, the total number of steps (defined as actions performed by a listener towards the solution to an equalization adjustment task and including mainly spectral adjustments and sound comparisons), and particularly the number of frequency adjustments and the number of
sound comparisons, were found to be most revealing of differences between the two groups.

Response time was found not to be by itself a significant and accurate performance indicator. This can be surprising at first. I had initially hypothesized that the ability to quickly identify the location of spectral resonances was a good indicator of a robust and well developed memory of timbre references. A listener who does not know where a resonance is in the spectrum will have to try more possible frequencies than another listener who knows. However, a listener can also provide an erroneous answer very quickly. An index of performance (PI) taking both accuracy and response time into account was then developed. Performance level is proportional to accuracy but decreases as response time increases. Subjects obtained PI values ranging from 41.4 to 123.0, with means of 55.8 for the control, untrained group, and 87.7 for the experimental, trained group. The experimental group, with little experience and 6 months of timbral ear training, thus outperformed the control group who had much more experience. These results suggest that the training as presented in this dissertation seems to accelerate greatly the process of developing key listening skills for the evaluation of timbre. Some control subjects did perform well, suggesting that the skills that are developed during the training and the tasks that are presented, are relevant to real-world audio professional work.
6.1.3 Monitoring problem solving in timbre/equalization adjustment tasks

The tutoring sessions scheduled during the training allow the instructor to observe listeners and ask them questions while they work on equalization adjustment tasks. Invaluable information can be gathered that way about the strategies and clues used by the listener to reach their answers. Tutoring also provides information about causes of errors but it has its limitations. Perceptual, subjective evaluations are often difficult to externalize and verbalize. In addition, listeners are not always aware of the cognitive processes involved during the solving of an equalization adjustment task and as a consequence have sometimes difficulty explaining why they made a particular step or why they chose a given frequency. A computer-based observation mechanism was developed to alleviate these problems and gather more precise information about listeners while they work. In this monitoring system, "sensors" are placed at interaction points in the software and send messages to an Observer object which is responsible for building and maintaining a context that describes the state of things during problem solving. This context contains representations of the problem to solve and the steps taken by the listener to reach an answer. Rules are used to group these individual steps into actions. Actions are in turn grouped into strategies. The mechanism can detect a number of errors and listening strategies. It also produces a trace of all interactions between the user and the
software during problem solving. As will be seen in the following section, the monitoring function carries great potential for further investigations.

6.2 Limitations and suggestions for further research

This final section looks at two important issues that are addressed in this dissertation, the evaluation of expertise level and the monitoring of problem solving, to identify limitations and to suggest potential avenues for further research. Other aspects of the training program that could benefit from improvements are also examined.

6.2.1 Assessing expertise

The performance index used in this dissertation certainly represents an improvement over the evaluation method that was used in Timbral Ear Trainer 1 and better reflects the skill level of the listeners by combining both accuracy and response time. However, the calculation of the index does not take into account the difficulty of each adjustment task. As a result, performance index of skilled listeners can vary greatly from one question to the other. For example, a skilled listener can recognize quickly that a given question has a single peak at 1 kHz and respond with the right adjustment very
quickly, obtaining a high PI value for this question. The same listener will understandably require more time to answer a more difficult problem in which there is a mixture of peaks and dips. With the current PI equation, the calculated value for the more difficult question will be lower. And hence, standard deviation for trained listeners tend to be higher than it is for untrained listeners. This is a difficult problem to tackle. How can the difficulty of a given question in equalization removing be assessed? How can the perceptual salience of the timbre differences between a given spectrally modified sound stimulus and the "flat" version of the same stimulus be evaluated? The perceptual difference between a sound on which a pattern of spectral peaks and dips was applied and the unprocessed version of the same stimulus depends on the initial spectral contents of that stimulus. It may also depend on the coupling of the room in which the training takes place and the loudspeakers that are used, which can produce room modes and resonances that can interfere with the perceptibility of fine, subtle timbral differences. Further work is clearly needed to address this issue.

Response time is believed to be a global indicator reflecting the number of adjustments, sound comparisons and listening time and hesitations of the listener in the course of solving a problem. A more precise and detailed assessment of a listener's abilities could perhaps be obtained by including in the calculation of the index secondary indicators such as the number of sound comparisons and spectral adjustments. Further research is also needed in that direction.
I have pointed out earlier that the measurement of response time is not very meaningful if the frequency adjustment is wrong. In the listening experiment reported in chapter 5, response time values were calculated for each question. However, many of the questions had multiple spectral modifications and listeners obtained partial scores, making both right and wrong adjustments within the same question. Perhaps this type of answers could be analyzed in more detail and the response time factor could be weighted differently based on the accuracy of each adjustment.

Further experimentation is finally needed in order to provide a scale of PI values that could indicate whether the expertise of a listener should be considered as low, medium, or high. Average PI values obtained by the subjects ranged from 41 to 123. In an informal experiment that I carried out on myself in which a set of questions having the same parameter ranges than those used in the listening test were generated by the computer, I obtained an average PI value of over 333, with values for individual questions ranging from 167 to 667. There is thus quite a large range of values that can be obtained and additional work is required to devise a scale on which the meaning of these PI values could be measured.
6.2.2 Problem solving monitoring

The monitoring mechanism described in this thesis can detect manipulation errors, infer basic problem solving actions and derive from those a few listening strategies. However, there are many sequences of actions that it is unable to group into relevant strategies, partly because of "noise" produced in the data, partly because of data generated by the user interactions that are unaccounted for by the rules that are defined. The system is also unable to learn new, unknown strategies. The extent of its usability in daily training thus has its limitations. It produces however trace files that describe in detail the sequence of interactions during the solving of each spectral adjustment problem. There is, as can be seen, room for improvement and many areas for further research and development of the monitoring system can be considered.

The context maintained by the monitoring system could be expanded to include intermediary results (e.g., in a multiple-band problem, it could indicate when each spectral modification is successfully adjusted and indicate also intermediate errors). The power spectrum of each sound stimulus should be added to the context as it can explain in part adjustments or errors a listener may make. Parsing of the trace files could be developed to extract sequences of actions that would show more clearly problem-solving phases and possible strategies. The results of the parsing of raw events from the trace files could be compared with an history of the context maintained by the Observer object.
during training. The results of these analyses could lead to a refined and more complete knowledge base that could then be used to provide better targeted feedback to the user.

The accuracy of the monitored events during training could be enhanced by providing more controlled and detailed problem solving operators. For example, the vowel operators as currently implemented do not prevent listeners from using them as shortcuts to quickly adjust frequencies in the vowel range even if instructed not to do so, and their use does not necessarily indicate that listeners did hear a resonance as a vowel. The interaction mechanism for this operator should be revised. The software could support coarser evaluation of center frequencies such as allowing the listener to indicate, for instance, if the spectral modification in a given band is located in the low, middle, or high frequency region. Feedback could be provided on the use of these operators.

I would like to point out here that production rules face their limitations when used in domains where knowledge is uncertain (such as the subjective evaluation of timbre differences) and behavior difficult to predict, since they must be explicitly encoded. However, there is enough procedural and declarative knowledge involved in a task such as equalization removing to make them useful as a monitor and feedback tool.
6.2.3 Training Program

The training has focused on developing memory and listening strategies for the evaluation of patterns of spectral peaks and dips in the spectrum of sounds. A complementary set of exercises could and should be developed to train listeners to evaluate the bandwidth of a sound's spectrum. Low-, high-, band-pass filters, and shelf-type equalization curves could be used to this end. Partial development of exercises concerned with high- and low-pass cutoff frequency adjustments has indicated that the skills currently trained in the course help greatly to solve this type of problem.

In sound recording practice, spectral adjustments are often applied in the context of several sound elements that need to blend together or segregate from others to achieve a desired balance. A natural extension of the training would be to allow independent processing of multiple sound elements (tracks) to implement, for example, exercises in which various mixes would have to be matched by adjusting level and equalization of the sound tracks. Although the basic listening skills developed in the training program are relevant to "real-world" audio work as the listening test results reported in chapter 5 suggest, this new type of exercise would provide more diverse and realistic contexts in which to apply the trained listening skills.

The seemingly endless and rapid increase of computing power available in desktop computers allow now the processing of sound in real time, without the need for
additional hardware. Using software-based DSP would allow much more flexibility in the type and specification of spectral adjustments that can be applied to the sound. For example, the order of audio filters could be adjusted, a function not available with the current hardware.

Much remains to be learned about the cognitive processes involved in timbre evaluation tasks such as described in this dissertation. It is hoped that the work presented here can suggest fruitful avenues along which research can be conducted, and thus contribute to demystify how listeners can become highly skilled in the evaluation of timbre. The ability to perform absolute identification of a set of timbre-based sound categories as trained in the course described above might perhaps be considered to be a form of acquired absolute timbre similar to absolute pitch. It would be valuable to pursue research on the relationship between the two skills.
The listening test today is divided into two parts. The first part will allow you to become familiar with the task procedure you will have to perform as well as to become comfortable with the software. You will be able to try a few questions until you feel comfortable enough to do the test.

The test consists in 15 questions. For each question, the sound you hear is processed by the computer using a parametric equalizer according to pre-defined specifications. Your task is to cancel the equalization applied by the computer so that the sound quality is restored to its original state.

The equalization specifications are the following:

Only the frequency and the gain parameters will change from one question to the other. The Q value is always the same and is set to 2.0. The possible frequency values are octaves ranging from 125 Hz to 16 kHz. There are two possible gain values: +6 dB or -6 dB. In each question, there will be one, or two, or three frequency bands that are modified by the computer. Any combination of frequency and gain values are possible. So if a
Appendix 1 – Listening Test Protocol

given question has three modified bands, they could all be set to -6 dB, or all to +6 dB or any combination of ±6 dB. You therefore will have to make 3 decisions for each question:

1) The number of bands that have been modified
2) For each modified band, whether it is a peak or a dip
3) For each modified band, the frequency value at which the gain setting is applied.

You can cancel the equalization applied by the computer by applying the opposite gain at the same frequency. For example, if a question consists of a 6dB peak at 2 kHz and a -6dB dip at 8 kHz, you must apply a 6-dB dip at 2 kHz and a 6-dB peak at 8 kHz to restore the original spectrum.

For each question, you can compare at will between sound B (the equalized version that you have to modify) and the Flat (non-equalized, original) sound which is the non-processed version of the sound. You are free to adjust the listening level to a comfortable setting. While the balance between B and Flat should be correct initially, you can adjust it if you believe they are not equal.

The sound example we use consists of a looped excerpt from a pop song. The sound will be playing continuously but you will hear it only while you are working on a question.
Appendix 1 – Listening Test Protocol

There will be no sound between questions. You can request the first question of the test by clicking on the "Start" button in the exercise window and the subsequent questions by clicking on the "Next" button in the same window. Each time you click on "Next", you will start by hearing the non-processed version of the sound ("Flat"). Using the space bar, you can toggle between B and Flat. You can hear the effects of your adjustments only when sound "B" is selected in the Mixer window. Once you think that your adjustments result in B and Flat sounding the same, you click on the "OK" button to register your answer. The sound will then be muted until you click on "Next" again.

There are two ways to adjust the frequency and gain in the equalizer window. You can use the frequency fader and move its knob to the frequency that you think has been modified. When you release the fader's knob, the fader will be positioned at the closest octave value. If you know how to associate vowels with timbre, you can also adjust the frequency value by using one of the 5 vowel buttons in each band. If you don't know, please use the frequency fader. The fader will then be adjusted to the corresponding frequency value within the vowel frequency range. This implies that for lower and higher frequencies, you have to use the faders.

You can adjust the gain fader to the appropriate value by using the knob for coarse adjustments and the up/down arrows for finer adjustments. Or you can use the "P" button
Appendix 1 – Listening Test Protocol

to apply the peak value (6 dB), or the "D" button to apply the dip value (-6 dB) or the "-" button to reset the gain to 0 dB.

The "Reset Faders" button in the Equalizer window resets all equalization parameters and controls to the default, flat state.

The "Exercise Monitor" window contains basic progress information about the number of questions you have left to do and the parameter specifications for the test. You will not receive feedback about whether your answers are right or wrong, about your score, or about your response time.

For each question in the test, the goal for you is to adjust the equalization of sound B so that it sounds the same as the original "Flat" sound. Your answers are evaluated both on the accuracy of the adjustments and how quickly you can finish the problem (your response time). The response time is evaluated for each question separately and not for the whole test.

You are free to interrupt the test at any time if you need to by clicking on the Stop button. The test requires all your attention. If you need to take a pause to rest your ears, you can
Appendix 1 – Listening Test Protocol

do so between questions. If you do it while working on a question (before clicking on OK), the timer will keep running and your response time will be affected.

The test will automatically stop when you reach the last question and a message will indicate the end of the test in the Instructor window, at which time you should come see me.

If you have questions, I will be in the adjacent room. Otherwise, good luck and have fun.
Appendix 2 – Development Tools

1. Symantec C/C++ for Power Macintosh version 8.1

   C/C++ Compilers

   Symantec Corporation, 1996.

2. CLIPS (C Language Integrated Production System) version 6.1

   Originally developed by Gary Riley at the NASA Johnson Space Center. Now in public domain.


References


References


References


References


References


References


References


References


References


References


References


Sandell, G. J. (1997). "Definitions of the word timbre".

References

Friedman (Eds.), Handbook of perceptioncarterette (Vol. 2). New York:
Academic Press.

Shepard, R. N. (1963). "Analysis of proximities as a technique for the study of

210: 390-398.

Slawson, A. W. (1968). "Vowel quality and musical timbre as functions of spectrum
envelope and fundamental frequency," Journal of the Acoustical Society of
America, 43: 87-101.


Addison Wesley.

identify nonverbal sounds: An application of a computer as a teaching machine,"

345-361.
References


References


