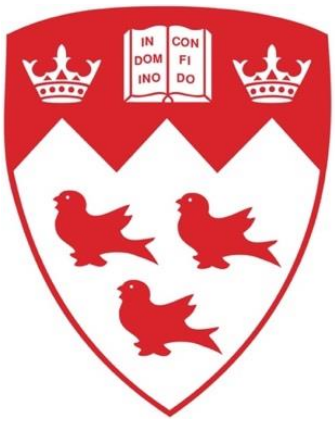
Toward Perceptual Searching of Room Impulse Response Libraries



David H. Benson

Graduate Program in Sound Recording

Department of Music Research

Schulich School of Music

McGill University, Montreal

August 5, 2022

A thesis submitted to McGill University in partial fulfilment of the requirements

for the degree of Doctor of Philosophy. © David H. Benson, 2022

ABSTRACT

Search interfaces for room impulse response libraries are critical to the operation of convolution reverb, a popular digital audio effect. Traditional designs for such interfaces, however, scale poorly to the large library sizes encountered in contemporary practice. This dissertation explores an emerging approach to search interface design with greater scaling potential, namely, an approach that organizes impulse responses (IRs) by their perceptual properties. Effective design of such perceptual search interfaces requires a detailed understanding of how IRs are perceived and how their properties can be efficiently visualized. This dissertation seeks to answer open questions in these areas.

Following a review of research on reverberation perception, an exploratory study was conducted focused on mixing engineering contexts. In the first part of the study, a set of putative reverberation attributes was identified by analyzing algorithmic reverb presets. Statistical models of these attributes were developed, and these models were used to probe the perceptual dimensions of a library of IRs measured in physical spaces. The study singled out two attributes for further investigation: reverberation brightness and sound source distance.

Next, a psychophysical experiment was conducted to identify the objective signal correlates of these attributes. Reverberation brightness was found to be better predicted by a novel signal feature, based on the spectral slope of the early portion of the IR, than by a set of candidate features drawn from the literature.

Finally, compact visual representations of the perceptual qualities of IRs were explored. A novel, task-specific visualization technique was developed. Compared with a generic visualization technique, the novel technique was associated with better performance on an IR library search task.

The novel visualization technique and brightness predictor have been incorporated into a prototype search interface for a vast library of high-resolution multichannel IRs. Future work will test the effectiveness of this perceptual interface in ecologically valid contexts.

RÉSUMÉ

Les interfaces de recherche pour les bibliothèques de réponses impulsionnelles de salles sont primordiales pour l'opération de la réverbération à convolution, un effet audio digital populaire. Pourtant, les conceptions traditionnelles de ces interfaces s'adaptent mal aux bibliothèques de tailles importantes rencontrées dans des usages contemporains. Cette thèse explore une approche émergente pour la conception d'interfaces de recherche avec un potentiel supérieur de mise à l'échelle, notamment une approche qui organise les RIs par leurs propriétés perceptuelles. Afin d'adéquatement concevoir ces interfaces de recherche perceptuelle, une compréhension précise de la manière dont les RI sont perçues et dont leur propriétés peuvent être visualisées efficacement est nécessaire. Cette dissertation vise à répondre aux questions ouvertes dans ces domaines.

Suite à l'examen de la recherche sur la perception de la réverbération, une étude d'exploration fut menée, focalisée sur le contexte du mixage sonore. Dans la première partie de l'étude, un groupe d'attributs putatifs fut identifié par l'analyse des réglages prédéfinis de réverbérations algorithmiques. Des modèles statistiques de ces attributs furent développés, et ces modèles furent utilisés afin de sonder les dimensions perceptuelles d'une bibliothèque de RIs mesurée dans des espaces physiques. Deux attributs furent distingués par l'étude pour être analysés de manière approfondie: la brillance de la réverbération et la distance de la source sonore.

Ensuite, une expérience psychologique fut menée afin d'identifier des corrélations entre ces attributs et des caractéristiques objectives des signaux. La brillance de la réverbération s'est avérée être mieux prédite par une caractéristique du signal novatrice, basée sur la pente spectrale de la première partie de la RI, plutôt que par un ensemble de caractéristiques candidates tirées de la littérature.

Enfin, des représentations visuelles compactes des qualités perceptuelles des RIs furent explorées. Une nouvelle technique de visualisation spécifique à certaines tâches fut développée. Comparée à une technique de visualisation générique, la nouvelle technique fut associée à de meilleures performances lors d'une tâche de recherche dans une bibliothèque de RIs.

La nouvelle technique de visualisation et la prédiction de la brillance ont été incorporées dans un prototype d'interface de recherche pour une vaste bibliothèque de RIs de haute résolution aux canaux multiples. À l'avenir, les travaux évalueront l'efficacité de l'interface perceptuelle dans des contextes écologiquement valables.

ACKNOWLEDGEMENTS

This work would not have been possible without the guidance of Prof. Wieslaw Woszczyk, whose fascination with reverberation inspired the research questions explored herein. Thanks are also due to Prof. William L. Martens who encouraged the project in its earliest stages, and to Prof. Richard King who provided invaluable feedback toward its end.

I wish to thank my colleagues in the sound recording department for their camaraderie and constructive criticism, especially at our monthly colloquia: Aybar Aydin, Alejandro Aspinwall, Vlad Baran, Matt Boerum, Eric Gaskell, Gianluca Grazioli, Johnathan Hong, Will Howie, Sungyoung Kim, Jack Kelly, Doyuen Ko, Brett Leonard, Bryan Martin, Denis Martin, Greg Sikora, Jamie Tagg, Diego Quiroz, Kent Walker, and Ying-Ying Zhang.

In addition to my colleagues in the PhD program, a special mention must also be given to my collaborator Vanille Patier-Debray, who made essential contributions as abstract translator, software co-developer, and IR library manager.

I acknowledge funding for this research from the AES Educational Foundation (John Eargle Award), the Fonds Quebecois de Recherche sur la Société et Culture (FQRSC), the James McGill professorship, the National Science and Engineering Research Council (NSERC), the Social Sciences and Humanities Research Council (SSHRC), and the Center for Interdisciplinary Research in Music, Media and Technology (CIRMMT).

A further debt is owed to the musical organizations I've worked with over the course of my degree: the chorus of the Montreal Symphony Orchestra, the choir of the Church of St. Andrew and St. Paul, One Equall Musick, and the Liederwölfe Opera Collective, among many others. It was music that first led me to audio and acoustics, and keeping one foot in the musical community during my graduate studies has been crucial for my sanity.

Lastly, this dissertation would not have been possible without the support of my family. I thank my father Jim for his counsel on navigating academia, my brother Chris for his statistical advice, my wife Joanna for her editing skills and infinite patience, and my daughter Saskia, for the joy she brings me.

To paraphrase the great comedian Maria Bamford, I owe this work to the blessings I've been chanced with: a modicum of talent... and every possible advantage.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

1 A RATIONALE FOR PERCEPTUAL SEARCH INTERFACES

Artificial reverberation is a category of popular signal processing effects widely applied in mixing and audio production contexts. One method of generating artificial reverberation, that of convolving a dry source signal with a measured room impulse response (IR), is lauded for the naturalness of its output and its ease of use. Reverberation generated in this way is known as convolution reverb. Manipulating convolution reverb generally requires searching through libraries of impulse responses. The sizes of these libraries have grown tremendously in recent years, such that existing search interface designs are no longer well adapted to them.

Motivated by one especially large collection, the Spacebuilder IR library (see Appendix A), this dissertation explores an approach to search interface design centred around the perceptual properties of IRs. This perceptual approach holds promise for the efficient navigation of large collections. Designing an effective perceptual search interface for IR libraries, however, would require a detailed understanding of how reverberation is perceived in mixing engineering contexts. The primary aim of this document is to advance knowledge in this area.

This introduction section will further motivate perceptual IR search interfaces by discussing current approaches to convolution reverb user interface design, and by contrasting these designs with those used in other types of reverb effects. The basic structure of a perceptual search interface will be outlined, and a key concept used throughout this work will be explained: that of a "perceptual model" of reverberation. Finally, the structure of the dissertation will be presented.

1.1 Studio reverberation effects: algorithmic vs convolutional approaches

Two principal methods exist to create artificial reverberation in audio production pipelines. These might be called the algorithmic and convolutional approaches. Although this work is focused on convolution, user interfaces associated with algorithmic reverb have instructive properties. Below, these two approaches are discussed, focusing on differences in their user interfaces. Some shortcomings of standard convolution reverb interfaces are identified, and remedies for these shortcomings that borrow design elements from algorithmic reverb interfaces are presented.

1.1.1 Control interfaces for algorithmic reverb

In the algorithmic paradigm, a digital algorithm simulates the effect of an acoustical enclosure on an input signal. Reverberation algorithm designs are reviewed by Gardner (2002) and Välimäki et al. (2012, 2016). The behaviour of the algorithm is parameterized by a number of control values, some of which are typically exposed through a user interface. Two examples of algorithmic reverb UIs are shown in Figure 1 and Figure 2. The first example includes parameters named "reverberance", "heaviness" and "liveness", among others, and the second includes parameters named "reverb time", "diffusion" and "tail width", among others.

Although the names of algorithmic control parameters vary greatly across plugins (Garland & Ronan, 2021), they tend to share a common property of manipulating perceptual attributes of the resulting reverb in predictable ways. That is, varying a control parameter named "diffusion" would tend to modify the output signal such that an experienced sound engineer would describe it as "more diffuse", while ideally not changing any other aspects of the reverb besides its diffusion. Algorithm designers generally strive for a one-to-one mapping between orthogonal perceptual attributes and control parameters, although this is difficult to achieve as many aspects of reverberation perception remain matters of scholarly debate (Gardner, 2002).

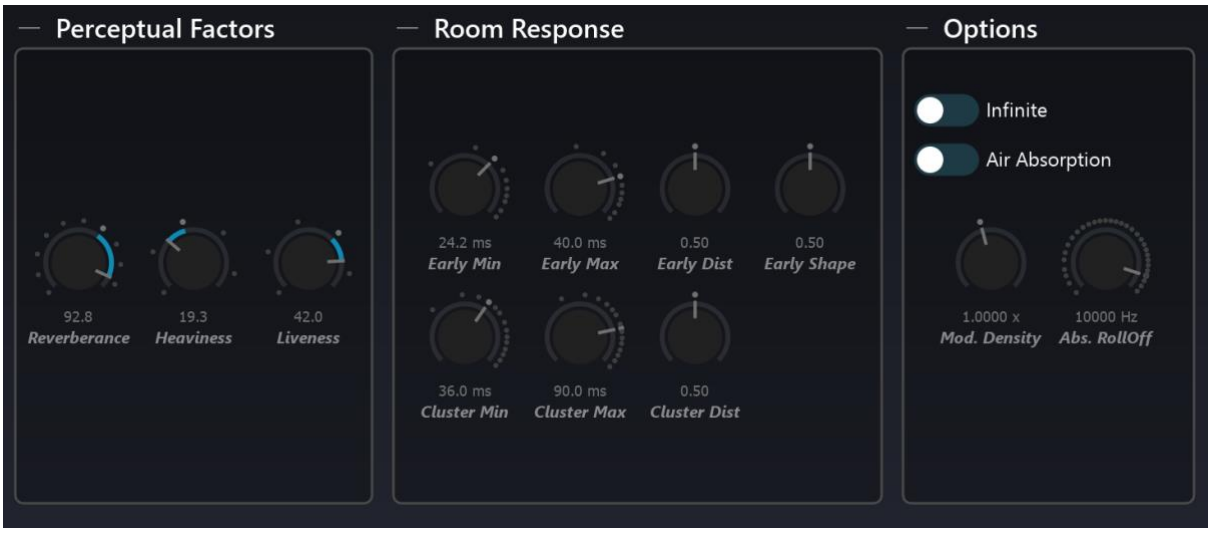


Figure 1 Spat Revolution control parameters[1]

1.1.1.1 Algorithmic reverb presets

In practice, using algorithmic reverb typically involves iteratively adjusting control parameters until a perceptually appropriate output has been found. As an alternative to adjusting parameters one by one, user interfaces may also allow users to load saved parameter configurations called *presets*. Loading presets can drastically change the sound of the resulting reverb as many parameters are modified at once. Presets are often given names that describe their perceptual character, as will be examined in chapter 3.

[1] Screenshot, by the author, of Spat Revolution (Flux:: Software Engineering, 2021).

Figure 2 Lexicon PCM Native control parameters[2]

[2] Screenshot of the PCM Native Reverb plugin (Lexicon, 2021) by Jack Kelly. Used with permission.

1.1.2 Control interfaces for convolution reverb

Convolution reverb effects generally differ from algorithmic effects both in their internal signal processing methods and in their user interfaces. Internally, convolution effects function not by passing a dry input signal through a parameterized algorithm, but rather by convolving an input signal with an acoustical measurement called a room impulse response (IR). Convolution is a simple mathematical operation, equivalent to a multiplication in the frequency domain, that effectively imparts the acoustical properties of the IR onto the input signal. This results in an output signal containing the musical material of the input and the spatial and room-related characteristics captured by the IR.[3]

Like algorithmic interfaces, convolution reverb user interfaces may also expose perceptual control parameters. Rather than manipulating an algorithm, however, these parameters typically modify the currently selected IR in perceptual relevant ways, e.g., changing the inter-channel correlation to modify perceived spaciousness (Ben-Hador & Neoran, 2004; Sierra, 2018). As these control parameters manipulate an existing acoustic measurement, rather than adjust a flexible reverberation algorithm, they are typically restricted in the degree of perceptual variation they can affect.

Even when parametric controls are available, the principal technique for adjusting the character of convolution reverb consists of selecting and loading new IRs from the user's library. This selection process requires some method of navigating the library. If user's IR collection is small and familiar, the selection process can be quick and effortless. With larger or more novel collections, however, the task of locating a new

[3] Although, in the context of this work, impulses responses are assumed to capture the reverberation of natural spaces, in general, IRs can have varied origins. IRs characterizing the transfer functions of signal processing devices, for example, can be created using methods similar to those for physical spaces (Farina, 2000). IRs can also be synthesized, using so-called room acoustic modeling techniques, from three-dimensional digital models of architectural spaces (Brinkmann et al., 2019; Savioja & Svensson, 2015). Finally, novel IRs can also be produced via signal-based analysis and resynthesis techniques. Such techniques are particularly useful for correcting measurements that are defective in some way, due, for instance, to imperfect excitation signals or high noise floors (Abel et al., 2010; Bryan & Abel, 2010).

IR with a particular set of perceptual qualities can be daunting, and is tedious enough to slow down a production workflow. As audio pedagogue Alex Case explains:

> With limited adjustability, convolution can back the engineer into a signal-processing corner. The sound is perfect, almost. No small tweak is available that can bring it in line with what the engineer really needs. On a nonconvolution reverb, engineers simply adjust the appropriate parameter. On a convolution reverb, engineers most commonly hunt around for another impulse response and hope it sounds similar, only shorter. That is a different process. Hunting for different impulse responses and auditioning them is clumsy enough to interfere with the creative process (2012, p. 297).

The central assumption motivating this research is that the "clumsiness" Case attributes to IR searches results directly from deficits in existing IR search interface designs. In addition, it assumes that these deficits are related to the types of metadata the designs feature.

In general, user interfaces for searching digital libraries tend to make visible and searchable certain types of information that describe the collection's items. This data describing items in the collection is known as metadata. Metadata in an interface for browsing photo libraries might include date and location information, while metadata for browsing recorded music libraries might include artist and album names.

In current interfaces for searching IR libraries, metadata tends to describe the process by which the IR was measured: for example, the name of the venue in which the measurement was made, or the distance between the speakers and microphones used in the measurement. Examples of such "how and where" metadata can be seen in the search interface of Audio Ease's Altiverb, a prominent commercial convolution reverb plugin. Altiverb's search interface, shown in Figure 3, allows venues to be filtered by keyword or category and ordered by size or name. Photos of the venues are shown. Once a venue is selected, the IRs measured in it are listed according to source-receiver distance.

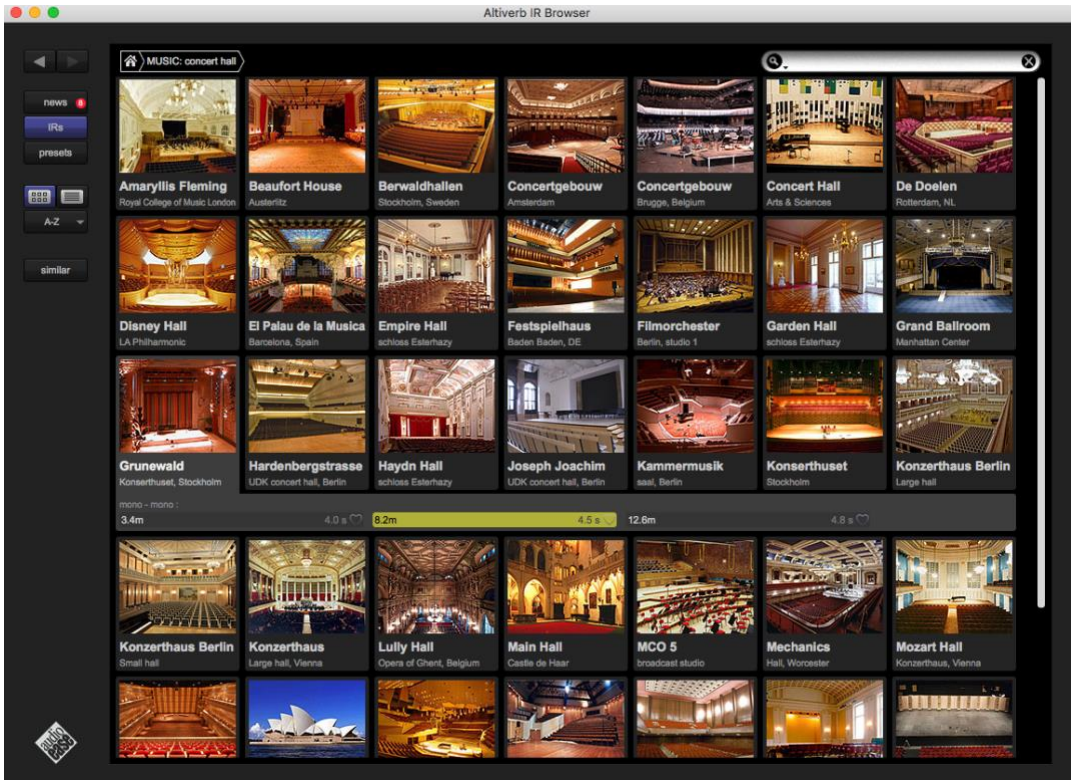


Figure 3 Altiverb IR browser[4]

This metadata gives a detailed account of each IR's origin and also provides some indirect information about its perceptual qualities, including its subjective decay time, timbre, and the perceived distance of sound sources convolved with it. These perceptual qualities are implicitly suggested by metadata items such as the venue size, venue photos, and physical source-receiver distances. Venue size is correlated with subjective decay time; venue photos provide information about surface materials and frequency-dependent absorption; physical source-receiver distance is related to perceived auditory

[4] Screenshot of the Altiverb convolution reverb plugin (Audio Ease, 2012) by Jack Kelly. Used with permission.

source distance. In each of these cases, though, it should be stressed that the psychoacoustic information conveyed by the metadata is ambiguous and imprecise. Small venues may have surprisingly long decay times (Ramakrishnan & Grewal, 2008), and physical source distance is only loosely related to perceived auditory source distance, even in controlled environment (Zahorik et al., 2005). Furthermore, in addition to their ambiguity, venue photos are also inefficient conduits for psychoacoustic data: at a legible size, only a small number of photos can be displayed simultaneously.

In summary, then, the central method for adjusting the character of convolution reverb is through selecting novel IRs from within a library. Although current library search interfaces may provide rich details about IR measurement processes, the perceptual information they present is often difficult to interpret. This lack of easy-to-interpret perceptual metadata makes it difficult to locate IRs with particular perceptual characteristics.

1.1.3 Perceptual search interfaces for convolution reverb IRs

A key difference between user interfaces for algorithmic and convolution reverb, then, is the ease with which individual reverberation attributes can be adjusted. Well-designed algorithmic reverbs exhibit a tight coupling between perceptual attributes and control parameters, allowing attributes to be easily manipulated independently. Independent attribute adjustment is more difficult in convolution reverb. Here, the primary method of adjusting reverb is through IR library browsing, and libraries are neither presented in ways that make attribute values clear, nor in ways that enable perceptually targeted searches (e.g., finding an IR like the current one but brighter). More succinctly, algorithmic reverbs have perceptual controls, but convolution reverbs do not enable perceptual browsing.

This research is premised on the idea that enabling perceptual browsing would make convolution reverb effects more efficient, and decrease the time spent on IR library searches in audio production workflows. Designing an effective perceptual browsing interface, however, would require a detailed understanding of reverberation perception,

since the interface's metadata would need to capture the relevant perceptual characteristics of IRs. A mathematical construct encapsulating a set of perceptual characteristics will be referred to in this dissertation as a "perceptual model". Perceptual models of reverberation are discussed in the next section.

1.2 Perceptual models of room reverberation

In the context of this thesis, the term *perceptual model* will be used to refer to a multidimensional space in which objects can be positioned such that their proximity reflects their perceived similarity. That is, objects judged to be similar will be close together in the space, while objects judged to be dissimilar will be further apart.

Models of this sort have long been used to conceptualize multidimensional perceptual constructs such as musical timbre (e.g., Grey, 1977). An example of such a model in shown in Figure 4. This image shows a two-dimensional space in which instrument sounds are positioned. Similar sounding wind instruments such as the flute and baroque recorder appear close together on the top left of the image, while similar sounding keyboard instruments, such as the piano and harpsichord appear at the right (Lakatos, 2000).

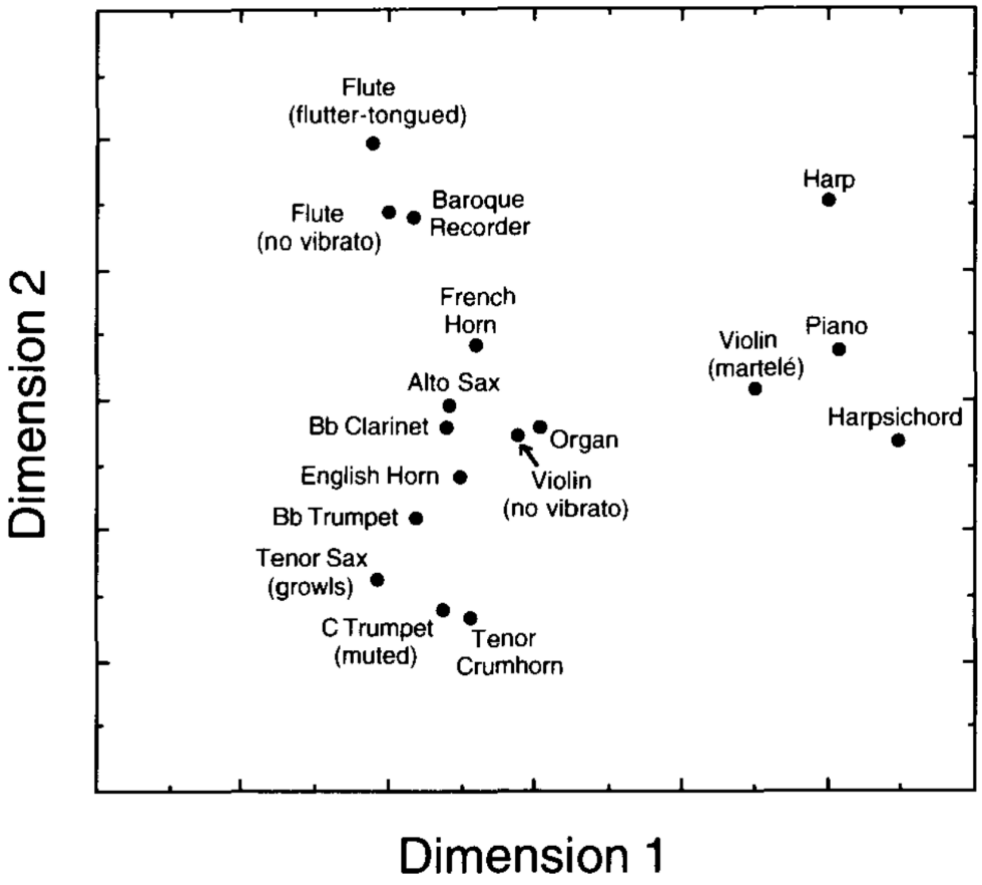


Figure 4 A perceptual model of musical instrument timbre[5]

[5] Reprinted by permission from Springer Nature *Perception and Psychophysics* (A common perceptual space for harmonic and percussive timbres, S. Lakatos), copyright 2000.

1.2.1 Attributes

A perceptual model is defined by one or more dimensions, each of which corresponds to a perceptual attribute that can vary independently. In the example, the first dimension characterizes the attack of the instrument: instruments with sharp attacks resulting from plucked or hammered strings (e.g., harpsichord, piano) appear at right, while instrument with smoother attacks (brass, reed, and wind instruments) appear at left. The second dimension characterizes the instruments' frequency content, with brilliant, muted trumpet appearing at one extreme and more rounded flute sounds appearing at the other. The dimensions of a perceptual model are sometimes given descriptive names: here, the first dimension/attribute might be named "attack smoothness" and the second dimension/attribute might be named "brightness".

Although models with only two dimensions are easiest to depict visually, in general, an arbitrary number of dimensions are possible. Three-dimensional models are commonly referenced in musical timbre research (e.g., McAdams, 2019)), and still higher-dimensional models are needed to describe phenomena that can vary along many independent perceptual attributes. As will be discussed in later chapters, room reverberation would appear to be one such higher dimensional phenomenon.

1.2.2 Objective signal features

A second component of a perceptual model, in the sense the term is used in this dissertation, is a set of physical properties of sound signals that correlate with the model's dimensions. In the timbre example above, the first dimension happens to correlate strongly with the time between a tone's onset and its maximum, the so-called "attack time", and the second dimension happens to correlate strongly with the centre of gravity of the tone's magnitude spectrum, or the so-called "spectral centroid". These objective signal features are useful because, if they are sufficiently well correlated with dimensions of the model, they can be used to predict the values of an objects' perceptual attributes. In an ideal case, for instance, objective "attack time" would precisely predict perceived "attack smoothness" ratings, and "spectral centroid" would precisely predict perceived "brightness" ratings.

In summary then, in the context of this dissertation, a perceptual model is a construct used to characterize perceived similarities and differences between a group of objects. A model has two components: a set of attributes on which objects can be compared (e.g., "attack smoothness", "brightness"), and a set of objective signal features that correlate strongly with model dimensions and can predict attribute values (e.g. "attack time", "spectral centroid").

Perceptual models are relevant to the challenge of controlling convolution reverb effects because they can inform metadata design. If an accurate perceptual model of room reverberation could be determined, its attributes could be used to structure the search interface's metadata: one metadata field could correspond to each of the model's attributes. This would ensure that the metadata contained a concise description of the perceptual character of the library's IRs, rather than information of weaker perceptual relevance. Further, the objective signal features in the model could be used to automatically assign values to each perceptual attribute. If perceptual attribute values could be predicted computationally for each IR, this would allow accurate metadata to be generated for arbitrarily large IR collections, potentially allowing the search interface to scale well to large libraries.

A complete specification of a perceptual model of room reverberation, then, would include a set of perceptual attributes and a set of objective signal features correlated with these attributes. Such a model could be productively used to inform the design of an IR search interface, much in the way that perceptual models of musical timbre have been used to control digital sound synthesis (e.g., Wessel, 1979). The optimal structure of room reverberation models remains a subject of scholarly debate, however. Considerable effort in the pages to follow will be spent developing a perceptual model that describes the attributes of convolution reverb impulse responses as perceived by mixing engineers, specifically.

1.3 Perceptual browsing via dynamic queries

While defining a perceptual model of room reverberation is a crucial step toward a perceptual search interface for IRs, it does not solve the design problem completely. Essentially, a perceptual model, complete with attributes and signal features, serves only to map each IR to a point in a perceptually meaningful multidimensional space. Associating a small number of attribute values with each IR then moves the design problem into a different domain: that of searching within large collections of multidimensional data. Fortunately, this more general problem has been explored in some detail in the human-computer interaction literature and some effective general solutions have been found.

One of the most enduring user interface designs for browsing multidimensional data is the so-called "dynamic query" interface (Ahlberg et al., 1992). In essence, dynamic query interfaces are 2D scatterplots with sets of filter widgets, one for each data dimension. The term "dynamic" here refers to a tight coupling between filter actions and visualized search results: as filters are adjusted, results matching the given query should appear nearly instantaneously in the display. Early applications of the paradigm involved browsing interfaces for real estate listings, where data dimensions included price, number of bedrooms and square footage (Ahlberg & Shneiderman, 2003), and for film libraries, where data dimensions included length and genre (Ahlberg & Shneiderman, 1994).

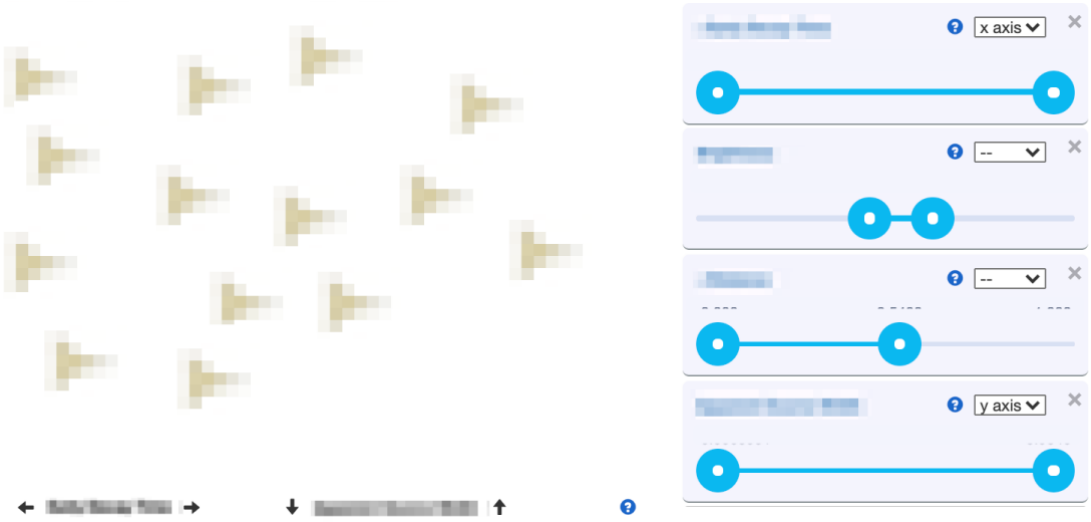


Figure 5 A dynamic query interface for IR library browsing

To give a concrete example, a mock-up of a dynamic query interface for IR library browsing is shown in Figure 5. Some portions of the figure are unclear because they depend on research questions to be answered later. For example, the figure shows filters for four data dimensions / perceptual attributes. What should these attributes be named, how many should there be, and to which objective signal features should they correspond? These questions depend on the perceptual model of reverberation chosen for the interface. Details aside, the figure aims to illustrate how such a search interface might function. The range filters at right would allow users to select ranges of perceptual attribute values, and IRs within these ranges would be shown in the display area. Clicking on an IR's icon might load it into a convolution engine to be auditioned.

A second area of the figure which is unclear is display section to the left, where query results are shown. How should IRs that match dynamic filter queries be drawn in this section? Early research displayed dynamic query results simply as dots or coloured bars, but noted that interface usability might be improved though more sophisticated mappings between data dimensions and visual dimensions of display icons (Ahlberg & Shneiderman, 2003). For example, an IR search interface might be improved by icons

that depicted an IR's perceptual properties. In general, a display icon that depicts the underlying values of a multidimensional data point is called a *glyph*. Perceptually-informed IR glyph designs are relatively unexplored research topic, yet an effective glyph design might be a useful component of a convolution reverb-centric dynamic query interface. The question of whether glyph design choices can improve IR search efficiency is relevant to the topic of search interface design and will be addressed later.

1.4 Dissertation outline

The goal of this dissertation is to lay the epistemic groundwork for the design of perceptually-oriented browsing interfaces for convolution reverb impulse response libraries. In particular, it aims to determine a useful perceptual model for convolution reverb IRs that could be used as a basis for a metadata scheme, and as a basis for a dynamic query-style search interface similar to that discussed above. The work also aims to determine whether perceptually-informed IR glyph designs can facilitate convolution reverb IR searches.

Throughout this work, one particular IR collection will be used as a case study for developing perceptual models and visualizations. This is the Spacebuilder IR Library, described in detail in Appendix A. Assembled over the past decade as a resource for high quality multichannel reverberation, the Spacebuilder library includes measurements at multiple source and receiver positions in over 200 venues, and, in total, includes over 10,000 multichannel IRs. Unless stated otherwise, all stimuli used in subjective experiments in this work incorporate IRs from this library.

The first part of the dissertation, Chapters 2 though 4, will be broadly focused on investigating the perceptual structure of room reverberation. Chapter 2 will review the existing literature on reverberation perception and will identify a promising preliminary model for convolution reverb IR libraries. Chapter 3 will explore refining this preliminary model by complementing it with additional attributes. A set of putative additional attributes will be identified from an objective analysis of algorithmic reverb presets; these putative attributes will then be evaluated in a subjective experiment. The results of this experiment will single out two attributes for further investigation: sound source distance and reverberation brightness. Chapter 4 will consider objective signal feature correlates for these attributes via a second experiment.

Finally, Chapter 5 will turn its attention away for perceptual models and toward IR visualization techniques. A novel IR glyph design will be proposed that aims to facilitate searching within large libraries. This novel design will then be evaluated in a controlled experiment.

Chapter 6 will review the key contributions of the dissertation and suggest directions for future study.

2 PERCEPTUAL MODELS OF REVERBERATION

As stated in the previous chapter, a crucial component of a perceptual search interface for room impulse responses is a perceptual model of room reverberation. An interface that allows searching along perceptual attributes must make use of attributes that are both salient in the minds of users and that vary within the collection being searched. Furthermore, the attributes must be named and described in a way that is meaningful to users, and values for these attributes must be accurately assigned. As before, this thesis will use the term *perceptual model* to refer to both a set of attributes, or dimensions of perceptual variation, and a set of signal features that predict these attributes.

This chapter will examine the existing literature on perceptual models of natural reverberation with the aim of identifying a model that shows promise for use in an IR search interface. The qualifier “natural” here is used to specify reverberation that has been produced by a physical space, and possibly captured in an IR. It contrasts with the term "synthetic", used here to describe reverberation produced via a digital algorithm. The perceptual spaces of both types of reverberation almost certainly differ; this chapter will be focused on the former. Throughout, particular attention will be paid to one model of reverberation which is frequently cited in the research literature: that presented in annex A of ISO standard 3382-1 (International Organization for Standardization, 2009). Due to its inclusion in an ISO standard, this model is assumed to represent a consensus view within this acoustics community, at the time of its publication, of the most salient attributes of concert hall reverberation, and of the most effective signal-based methods of predicting them.

The chapter will be divided into three sections. First, it will review key findings from the century leading up to the publication of the Standard in 2009; it will then present the ISO

model and show how it incorporates these prior findings. The second section will focus on research conducted since 2009. Exploratory studies aimed at identifying attributes of reverberation will be treated first, followed by more focused studies aimed at improving prediction techniques for particular attributes. In a final concluding section, the relevance of the reviewed findings to IR search interfaces will be considered, and relevant open questions in acoustics perception will be discussed.

2.1 Perceptual models prior to 2009

Much contemporary research into room acoustics perception draws heavily on a perceptual model presented in ISO standard 3382-1, titled *Acoustics – Measurement of room acoustic parameters – Part 1: Performance spaces* (International Organization for Standardization, 2009). The aim of this section is to describe this model and position it in a historical context. To do so, key investigations into acoustics perception from the last century will be reviewed. Each investigation's methods will be discussed and the perceptual attributes it identified will be listed. At the end of the section, the ISO 3382 model will be presented and contrasted with the studies preceding it.

An overview of the results summarized in this section is given in Table 1. Columns of the table correspond to investigations and rows correspond to perceptual attribute names, grouped by semantic similarity. Only descriptive attributes are shown; hedonic attributes, such as *preference*, are excluded from the table. Additionally, when an attribute was shown to correlate strongly with an objective signal feature similar to one in the ISO 3382 model, that feature is shown in parentheses. These objective signal features will be defined later, in section 2.1.2.

The review begins with the investigations of Wallace Clement Sabine in Cambridge, MA, conducted near the turn of the 19th century.

Table 1 Summary of perceptual models pre-2009

Attribute category	*Sabine (1900)*	*Beranek (1962)*	*Hawkes & Douglas (1971)*	*Wilkens (1977)*	*Barron (1988)*	*ISO 3382-1 (2009)*
Loudness-like	Loudness	Loudness of direct sound		Strength and extension of the sound source	Loudness	Subjective level of sound (G)
Reverberance-like		Liveness (T) Loudness of reverberant sound	Reverberance (T)		Reverberance (EDT)	Perceived reverberance (EDT)
Clarity-like	Clarity		Definition	Clearness	Clarity (EDT)	Perceived clarity of sound (C80)
ASW-like						Apparent source width (IACC, J_{LF})
LEV-like		Diffusion			Envelopment (J_{LF})	Listener envelopment (L_J)
Timbral	Timbral balance	Warmth	Brilliance	Timbre	Treble-to-mid-frequency balance Bass-to-mid-frequency balance	
Misc		Intimacy Balance/blend Ensemble	Intimacy Evenness		Intimacy	

2.1.1.1 Sabine (1900)

Modern research into room acoustics perception began with the work of Wallace Clement Sabine at Harvard University, who, as a young physics professor, was asked to improve speech intelligibility in the lecture hall of the recently constructed Fogg Art Museum. According to his reports, before his interventions, speech sounds remained audible in the hall for over 5 seconds. At normal speaking rates, this meant that each sound interfered acoustically with 15-30 subsequent syllables. By comparing the defective hall to architecturally similar spaces with superior acoustics, Sabine pinned the problem on a lack of acoustic absorption. Conditions in the hall were ultimately improved by adding more absorptive material (Sabine, 1900).

Sabine is perhaps best remembered for his formula relating a room's absorptive area and volume to its reverberation time, and for the unit of acoustic absorption that bears his name. On top of these contributions, however, he is also notable for having recognized the multidimensional nature of acoustics perception. In an essay from 1900, he articulates what might be considered room acoustics' first perceptual model:

> In order that hearing may be good in any auditorium, it is necessary that the sound should be **sufficiently loud**; that the simultaneous **components of a complex sound should maintain their proper relative intensities***;* and that the **successive sounds** in rapidly moving articulation, either of speech or music, **should be clear and distinct** [emphasis added], free from each other and from extraneous noises. These three are the necessary, as they are the entirely sufficient, conditions for good hearing. (p. 4)

In the text above, Sabine identifies three dimensions on which a hall can be acoustically deficient. To paraphrase, halls can be lacking in loudness, timbral balance, clarity, or some combination of the three. Restated in a way free of value judgements, this implies that halls exist in a three-dimensional perceptual space, where the three dimensions might be named *loudness*, *timbral balance*, and *clarity*.

With respect to objective predictors, Sabine's work only addressed the last of these attributes. He noted that clarity was inversely proportional to reverberation time, which in turn was determined by a room's volume and absorption, thereby hypothesizing a

relationship between the attribute and two quantifiable, objective properties of an enclosure. Nonetheless, as remarked by Kahle (1995), his conception of room acoustics as a multidimensional perceptual phenomenon was prescient, and foreshadowed many results in the century to follow.

2.1.1.2 Beranek (1962)

The next major figure to significantly advance the discipline was MIT professor and consultant Leo Beranek, who published his landmark *Music, Acoustics and Architecture* in 1962. Beranek's investigative methods differed from Sabine's in his collaboration with acoustical stakeholders outside of his discipline. Whereas Sabine's model appears to have been the product of discussions with his physicist colleagues and students, Beranek's model was developed through interviews and correspondence with experienced listeners from other professions, principally music critics and high-profile orchestra conductors.

Beranek's book presents a total of 18 room acoustic attributes, each of which is classified as either "positive" or "negative". The presence in a hall of positive attributes, such as *intimacy*, *warmth*, and *diffusion* contributes positively to the quality of the space, while the presence of negative attributes, such as *echo* and *noise*, detract from the space's quality. Further, Beranek subdivides his positive attributes into "dependent" and "independent" groups, proposing that the dependent ones can be predicted from combinations of the independent ones. Beranek's independent attributes are then assembled into a "rating scale of acoustic quality", such that a weighted sum of these attributes can be used to estimate the overall quality of a hall. This rating scale is calibrated to successfully predict the average hall quality judgements made by a set of 50 prominent musicians. Table 1 lists the eight positive independent attributes in Beranek's model.

2.1.1.3 Hawkes and Douglas (1971)

Like Sabine, the investigations of R. J. Hawkes and H. Douglas were also inspired by an acoustically deficient venue, in this case, London's Royal Festival Hall. Following

its 1951 opening, the hall received criticism of its poor acoustics, prompting the installation in the early 1960s of an electronic system for increasing its low-frequency reverberation time (Parkin & Morgan, 1965). Hawkes and Douglas' work was initially aimed at assessing the perceptual effects of this enhancement system.

Hawkes and Douglas used Beranek's model as a point of departure. In their studies, subjects attending concerts were asked to record their impressions of a hall's qualities on 16 attribute scales, these scales being a subset of the 18 presented in Beranek's book. The ratings collected on these scales were then submitted to a factor analysis to identify a smaller set of independent components that explained their data.

A factor analysis of ratings from four seating positions in four British concert halls yielded six independent factors. One factor was hedonic (enjoyment); the other five were descriptive. To these descriptive attributes, Hawkes and Douglas assigned the names *reverberance, balance and blend, intimacy, definition,* and *brilliance*.

They also examined correlations between these attributes and objective features computed from the halls' impulse responses. Differences in reverberance were associated with changes in reverberation time *T* (see section 2.2.2), differences in brilliance with high-frequency *T*, and differences in intimacy with the initial time delay gap, the time between the arrival of the direct sound from the stage and the first reflected wavefront.

2.1.1.4 Wilkins (1975)

One methodological weakness in Hawkes and Douglas' study was a lack of control over visual stimuli. Since subjects were physically seated in concert halls and could see their surroundings, it's possible that their attribute ratings were influenced not only by the sounds they heard but also their visual environments. A subject's rating of (acoustical) intimacy, for instance, could conceivably be biased by their physical distance from the stage and musicians.

This weakness was addressed by Wilkens, a doctoral student at the Technical University of Berlin.[6] Wilkens controlled visual variables in his experiments by capturing orchestral performances with a dummy head microphone and presenting recordings to his subjects over headphones. Except for its laboratory setting, his methods were similar to those of Hawkes and Douglas. He employed a long (19 attribute) questionnaire and a subsequent dimensionality reduction analysis, which in his case yielded a compact three-dimensional perceptual model. Wilkens' named his three attributes *strength and extension of the source*, *clearness*, and *timbre*. The timbre factor correlated with questionnaire terms such as soft/hard, brilliant/dull, rounded/pointed, and light/dark. He did not investigate correlations between his attributes and objective signal features.

2.1.1.5 Barron (1988)

By contrast, uncovering relationships between subjective attributes and impulse response-based signal features was a primary objective of Barron's study in 1988. Drawing on earlier research, he began with a set of eight attributes he presumed to be independent and then set out to find signal features that correlated with them. His attributes were assembled into a survey that was completed by expert listeners, mostly acoustical consultants, at public concerts. His subjects evaluated a total of 40 listening positions in 11 British concert halls.

Barron's statistical analysis confirmed that his eight attributes were largely uncorrelated. Further, he found that five of these attributes exhibited relationships with objective predictors, many of which would come to be adopted in ISO 3382-1. His loudness and intimacy attributes were both relatively well predicted by a feature he called *total sound level*, reverberance and clarity by the ISO *early decay time*, and envelopment by a weighted sum of the ISO *early lateral fraction* and the *total sound level*. On the other hand, his two timbral attributes, *treble-to-mid-frequency balance*, and *bass-to-mid-frequency* balance, were not found to correlate strongly with any of his

[6] The key findings of Wilkens' dissertation, written in German, were reported in English by Cremer and Müller (1982). The summary above is drawn from Cremer and Müller.

signal features. The physical correlates of his final attribute, *soloist-to-orchestra balance*, were not explored.

2.1.2 ISO 3382-1

ISO-3382-1 is a document produced by the technical subcommittee on building acoustics of the International Organization for Standardization that prescribes metrics for characterizing the acoustical quality of performance spaces. The document is largely concerned with methods for measuring a room's *reverberation time* (*T*), defined as the amount of time required for sound energy to decay by 60 dB after a sound source has been silenced.

The document's first annex, however, lists five "subjective aspects of the acoustical character" of auditoria, along with signal features that are associated with them. These five subjective attributes are *subjective level of sound*, *perceived reverberance*, *perceived clarity of sound*, *apparent source width (ASW)* and *listener envelopment (LEV)*. These attributes will hereafter be referred to as the five-factor ISO 3382 perceptual model, or more simply the five-factor ISO model. Additionally, in contexts where loudness differences do not exist between stimuli (i.e., in the experiments reported in Chapters 3, 4 and 5) references will be made to a four-factor ISO model which contains all five-factor attributes except for subjective level of sound.

Despite the influence of ISO 3382-1 on the academic community, the standard itself is not an academic document. Whereas academic documents place considerable emphasis on articulating the origins within the research literature of ideas they present, the standard is not bound by such norms. It gives no precise information about the origins of its five-dimensional model, claiming only that "there is reasonable agreement that [measures beyond reverberation time] are needed for a more complete evaluation of the acoustic quality of rooms" (p. v). Rather than simply assume that the ISO model represents a consensus view of the most salient dimension of room acoustics amongst acousticians and acoustics researchers, this section will briefly trace the lineage of each attribute in the model and show that each has some degree of experimental validation.

2.1.2.1 Subjective level of sound, perceived reverberance, and perceived clarity

Of the attributes in the five-factor model, four are familiar from the studies reviewed earlier in this chapter, as shown in Table 1. Indeed, attribute names that are semantically similar to the ISO attributes subjective level of sound, perceived reverberance and perceived clarity each appear in multiple studies. The relatively consistent presence of these attributes in exploratory research would seem to justify their inclusion in the ISO model.

2.1.2.2 Listener envelopment (LEV)

With respect to listener envelopment, similar attributes are present in two studies. This attribute shows up explicitly in Barron's model, and implicitly in Beranek's model, under the term "diffusion". Although the meaning of diffusion is somewhat ambiguous in a psychoacoustic context (see discussion in section 2.2.2.3), Beranek's description makes clear that he intends it to refer to an attribute that is spatial in nature (1962). In his words,

> [d]iffusion concerns the spatial orientation of the reverberant sound. The diffusion is best when the reverberant sound seems to arrive at the listener's ears from all directions in about equal amount. (p. 67)

The sense of sound arriving "from all directions in about equal amount" seems tantamount to the notion of listener envelopment present in the ISO standard. Thus, since Beranek and Barron both refer to listener envelopment-like attributes, this component of the ISO model would appear to be supported by at least some prior research as well.

2.1.2.3 Apparent source width (ASW)

The only ISO model attribute that does not appear in the reviewed studies is apparent source width. If ASW is indeed an important source of perceptual variation between

concert halls, what would explain its absence from so many early studies? One explanation may be a strong correlation in real halls between ASW and the subjective level of sound. Despite the phenomenological distinctness of the level and ASW percepts, level appears to be an important contributor to ASW: louder halls tend to have broader sound sources (Marshall & Barron, 2001). Since the types of statistical analyses employed by Wilkens, and Hawkes and Douglas are only able to identify factors that vary independently, these analyses would be unlikely to separate attributes with a high degree of correlation. Thus, these analyses might include only a single dimension in their outputs onto which the attributes of level and ASW would both be collapsed. Indeed, the results of Wilkens support this hypothesis, as one of his output dimensions, "strength and extension of the source", appears to reflect correlated variations in both attributes.

Rather than emerging from questionnaire-based in-situ research, the attribute of ASW appears to originate instead from theorizing and laboratory studies in the late 1960's. In 1967 Marshall hypothesized an attribute of performance spaces, unrelated to reverberance, that "generate[d] a sense of envelopment in the sound and of direct involvement with it" (Marshall, 1967). He named this attribute "spatial responsiveness" and noted that it was strongest in relatively narrow halls, where side-wall reflections arrived before ceiling reflections.

Later, in a laboratory setting, Barron tested Marshall's hypothesis that early-arriving lateral reflections contributed to a distinct spatial attribute. He found that even a single such reflection caused music to "gain body and fullness" and sound sources to "appear[…] to broaden" (Barron, 1971). He renamed the attribute "spatial impression" and noted that its magnitude was proportional to the strength of lateral reflections, so long as these arrived within 10 and 80 ms of the direct sound. While Barron's 1971 study investigated only single reflections, his later work confirmed the presence of the spatial impression attribute in more ecologically valid sound fields (Barron & Marshall, 1981). This latter paper also proposed the *early lateral energy fraction* as an objective measure to predict spatial impression. While Barron's objective measure made its way into the ISO standard, his name for the associated attribute did not. Rather than spatial impression, the ISO standard refers to this attribute as apparent source width, a term coined by Keet while carrying out similar research in parallel (1968).

The attribute of ASW, then, although largely absent from the five studies reviewed at the start of this chapter, is supported as a distinct attribute of room acoustics by multiple laboratory experiments.[7] The five-factor ISO model is summarized in the final column of Table 1. The signal features associated with the attributes of the model, shown in brackets, are described later, in section 2.2.2.

2.2 Recent research

Having summarized the perceptual model in ISO 3382 and the research that led to it, this chapter will now turn to reviewing significant results produced since the adoption of the 2009 standard. For the purposes of this discussion, recent research projects will be put into one of two categories based on their objectives.

In the first category are studies that seek to explore the perceptual space of reverberation. That is, they aim to identify the dimensions, or attributes, that explain perceived differences between reverberant sound signals. One theme uniting these studies is a lack of a priori assumptions about reverberation's perceptual structure. They begin with few preconceived notions about the number or character of attributes, and build perceptual models based on the experimental data collected.

In the second category are studies aimed at better understanding the physical correlates of subjective attributes: studies that seek to improve techniques for predicting attributes from objective signal features. Unlike studies in the first group, these studies do make strong assumptions about reverberation's perceptual structure. Namely, they assume that one or more specific attributes exist, and they then probe the relationships between these attributes and audio signals that prompt variations in them.

[7] An additional factor that may have contributed to the inclusion of ASW in the ISO model may be an association between it and hedonic hall quality. ASW is predicted by measures of interaural correlation, and interaural correlation measures, in turn, have been found to predict hedonic quality (Hidaka et al., 1995; M. R. Schroeder et al., 1974).

These two groups of studies will be treated in the following two sections. Exploratory studies will be dealt with first, followed by studies focused on the relationships between specific attributes and signal features.

2.2.1 Perceptual structure explorations

In recent years, important investigations into the perceptual structure of reverberation have been carried out at Aalto University in Finland and at the Technical University of Berlin. The methods employed at the two institutions are similar in both relying on laboratory sound reproduction rather than in-situ listening in concert halls. Their methods differ however in their use of provided constructs versus elicited constructs. Whereas Weinzierl et al., in Berlin, asked subjects to rate sound fields on a large set of provided attribute scales, Lokki et al., in Finland, first elicited unique attributes from individual subjects, and then asked subjects to make ratings on these elicited attributes. Lokki et al. cite Berg and Rumsey (2006) as an inspiration for applying individual attribute elicitation in a room acoustics context, as the pair had employed similar methods to evaluate spatial audio.

The key findings of these perceptual structure explorations are summarized in Table 2. As in the previous section, columns correspond to studies, and rows correspond to attributes identified in the studies, grouped by semantic category. For comparison, the five-factor ISO model is shown at left.

Table 2 Summary of perceptual models post-2009

Attribute category	*ISO 3382-1 (2009)*	*Lokki et al (2011)*	*Lokki et al (2012)*	*Weinzierl et al (2018)*
Loudness-like	Subjective level of sound (G)	Loudness/Distance		Strength
Reverberance-like	Perceived reverberance (EDT)	Reverberance related to size of the space Enveloping reverberance	Reverberance	Reverberance
Clarity-like	Perceived clarity of sound (C80)	Definition	Definition	
ASW-like	Apparent source width, ASW (IACC, J_{LF})	Width of sound		
LEV-like	Listener envelopment, LEV (L_J)		Envelopment / Loudness	
Timbral			Bassiness	Brilliance Colouration
Misc			Proximity	Irregular decay

2.2.1.1 Lokki et al. (2011)

The sound stimuli in both studies by Lokki et al. were captured by driving concert hall halls with a so-called “loudspeaker orchestra" (2011). The anechoically-recorded orchestra “played” four orchestral excerpts from the standard repertoire. Sound fields were measured at three positions in the audience area of three different halls. Four

musical excerpts by three halls by three receiver positions led to a total of 36 stimuli to be evaluated.

After familiarizing themselves with the stimuli, each subject was asked to generate a set of descriptive attributes that they felt could discriminate the stimuli. Subjects then rated each stimulus on each of these individual attribute scales. The ratings from the study's 20 subject, who produced 102 unique attributes, were then summarized using various dimensionality reduction techniques.

The researchers found five underlying attributes in the data, which they named *loudness/distance*, *width of sound*, *reverberance related to the size of the space*, *enveloping reverberance*, and *definition*. These attributes were then contrasted with the five-factor ISO model. Loudness/distance was successfully predicted by the ISO strength measure *G*,[8] and the width of sound attribute was well predicted by ISO late lateral level. The study's two reverberance-like attributes were only moderately-well predicted by the ISO's reverberance feature *early decay time (EDT)*. Finally, the study's definition attribute was not well predicted by any of the ISO's objective signal features, including the C_{80}, as might have been expected.

2.2.1.2 Lokki et al. (2012)

Subsequent work by Lokki et al. applied a similar experimental method to a larger set of nine concert halls. Here, the elicited attributes fell into five categories named *bassiness*, *envelopment/loudness*, *reverberance*, *definition* and *proximity*. In most cases, these attributes were reasonably well predicted by ISO features. Bassiness matched well with low frequency *G*, and envelopment/loudness correlated with the *early lateral fraction* and *late lateral level.* Reverberance and definition were moderately correlated with *EDT* and C_{80}. Curiously, though, the fifth attribute of proximity was found not to have a strong relationship with any ISO measure. This was a notable result, as proximity was the attribute most strongly correlated with subjective preference data, which was also collected.

[8] Precise definitions of ISO signal features such as *G, EDT,* and *late lateral level* are given in section 2.2.2.

2.2.1.3 Weinzierl et al. (2018)

In contrast to the individual vocabulary approach favoured by Lokki et al., Weinzierl et al. asked subjects to rate reproduced sound fields on a large set of provided attribute scales. The scales were created by a focus group of acousticians. Although the study's use of pre-defined scales was similar to the approaches of both Hawkes and Douglas and Wilkens, Weinzierl et al.'s dataset was much larger, including 190 subjects and 190 distinct stimuli (25 halls by two positions by three sound sources, plus 10 halls by two positions by two sound sources).

Owing to its scale, the study employed a balanced incomplete block design, where each subject rated only a subset of stimuli but where all stimuli were rated an equal number of times. A confirmatory factor analysis of the resulting data found a solution with five descriptive attributes.[9] The results were again contrasted with the five-factor ISO model, which yielded both similarities and differences. *Strength* and *reverberance* attributes appear in both, yet Weinzierl et al.'s model also contains three attributes with no ISO correlate, namely *irregular decay*, and two timbral attributes: *brilliance* and *coloration*.

Supplemental material included with the article explained which of the original set of 46 attribute scales were most strongly related to each of the five factors in the solution. Notably, the brilliance attribute was primarily related to scales describing the perceived amount of high-frequency energy in the stimulus, i.e. scales with poles labeled "dark/bright" and "attenuated/emphasized treble range". The coloration attribute was related to scales describing smoothness vs. unevenness in the frequency spectrum, e.g. the degree of "comb-filter coloration" and the degree of "boominess", where "boominess" was specified as a "low-frequency […] narrow-band resonance". The brilliance attribute has several clear antecedents in Table 1 and Table 2, while the coloration attribute is reminiscent of Sabine's description of timbral balance, in which "components of a complex sound should maintain their proper relative intensities".

[9] The solution also included one hedonic factor, *quality*, which is not discussed here in keeping with the chapter's focus on descriptive attributes.

Curiously, Weinzierl et al.'s model, at first glance, appears to lack any terms related to spatial impression. None of the five attributes identified in the study make any reference to width or envelopment. A close inspection of the study's results, however, affirms that subjects did indeed report differences in spatial characteristics between stimuli, but that these spatial attributes were correlated with non-spatial ones. Indeed, the model's strength attribute is highly correlated with a scale describing "the perceived spatial extent of a sound source in the horizontal direction" as well as a scale describing "loudness". Likewise, the model's reverberance attribute is highly correlated with a scale describing the degree of "envelopment by reverberation" as well as general "reverberance". These results reaffirm the notion, presented earlier in section 2.1.2, that the attributes of subjective level of sound and ASW tend to be highly correlated in natural acoustic spaces. They also suggest that the same might be true of reverberance and LEV to some degree.

2.2.1.4 Summary of post-2009 perceptual structure explorations

To briefly summarize this recent work, then, efforts by Lokki et al. have shown that research paradigms involving individually elicited attributes can be fruitfully applied in the concert hall acoustics. Their results generally reaffirm the ISO model, although the relatively small sizes of their stimulus and subject pools (3-9 halls; 17-20 subjects) temper the generalizability of their findings. Weinzierl et al.'s study is intriguing by virtue of its breadth and the contrasts between its results and the ISO model. Notably, his results highlight the interdependence of spatial and non-spatial acoustic attributes and the perceptual salience of two distinct timbral dimensions, there named brilliance and colouration.

2.2.2 Objective predictors

Having surveyed recent research into the broad perceptual structure of room reverberation, this chapter will now consider efforts to predict subjective values of attributes using objective signal features. The discussion will be split into four sections. First, the signal features proposed in ISO 3382-1 will be defined. Next, the chapter will

discuss efforts to predict two attributes not present in the standard: reverberation *timbre* and *diffusion*. Despite their absence from the standard, timbre and diffusion are two aspects of reverberation that are widely considered to be subjectively important. Finally, approaches to attribute prediction using binaural models will be considered. Binaural modeling approaches are fundamentally different from other techniques in the chapter because they operate not on impulse responses, but instead on *wet* signals, such as might be captured by dummy head during performances. Binaural modeling approaches have shown considerable promise in recent years, particularly with respect to predicting attributes from the ISO model. Impulse response-based features from the ISO 3382-1 standard will be treated first.

2.2.2.1 ISO 3382-1 signal features

The signal features proposed in ISO 3382-1 for predicting attributes of the ISO model appear in parentheses in Table 1 and Table 2 and have been referenced throughout this chapter. In this section these features will be precisely defined, and their associated attributes will be described. The discussion below is organized by perceptual attribute, first presenting signal features associated with perceived reverberance, then those related to subjective level of sound, perceived clarity, ASW and LEV.

2.2.2.1.1 Perceived reverberance (*EDT, T*)

Although the subjective experience of reverberance is not explicitly described in the standard, it is commonly understood to refer to the "subjective decay rate" of sound in a room (Lee & Cabrera, 2010). The standard recommends predicting perceived reverberance from an objective feature called the *early decay time (EDT).*

EDT is computed from a room's so-called energy decay curve (*E)*, typically computed using the integrated impulse response method of Schroeder (M. R. Schroeder, 1965). With this method, assuming that p is sound pressure, the decay curve E is defined as in Equation 1. E describes the amount of energy remaining in the impulse response at time t.

$$E(t) = \int_{t}^{\infty} p^2(\tau)\, d\tau$$

Equation 1 Energy Decay Curve

From E, the *EDT* is calculated from the line of best fit through the first 10 dB drop in the curve. The slope of this line of best fit, d, describes a rate of energy decay in decibels per second. EDT is then defined as $60/d$, which is the time required for an energy drop of 60 dB at decay rate d. The technique of predicting perceived reverberance from the initial portion of the decay curve, as opposed to a longer or later portion, was introduced by Atal, Schroeder and colleagues (Atal et al., 1965; M. Schroeder et al., 1966). The precise term "early decay time" was coined by Jordan (1970).

Although not the recommend predictor of perceived reverberance, the standard also gives methods of estimating a room's reverberation time, *T*. The derivation of *T* will be discussed here since it is relevant in later sections. Reverberation time is the amount of time required for sound energy to decay by 60 dB. Although *T* could, in principle, be calculated using the same method as EDT, but over a drop of 60 dB instead of 10 dB, in practice a shorter evaluation range is used to avoid complications related to background noise. A measurement called the T_{30} is a common estimate of *T*, where the subscript indicates an estimation using the 30 dB drop between the -5 and -35 dB points on the energy decay curve.

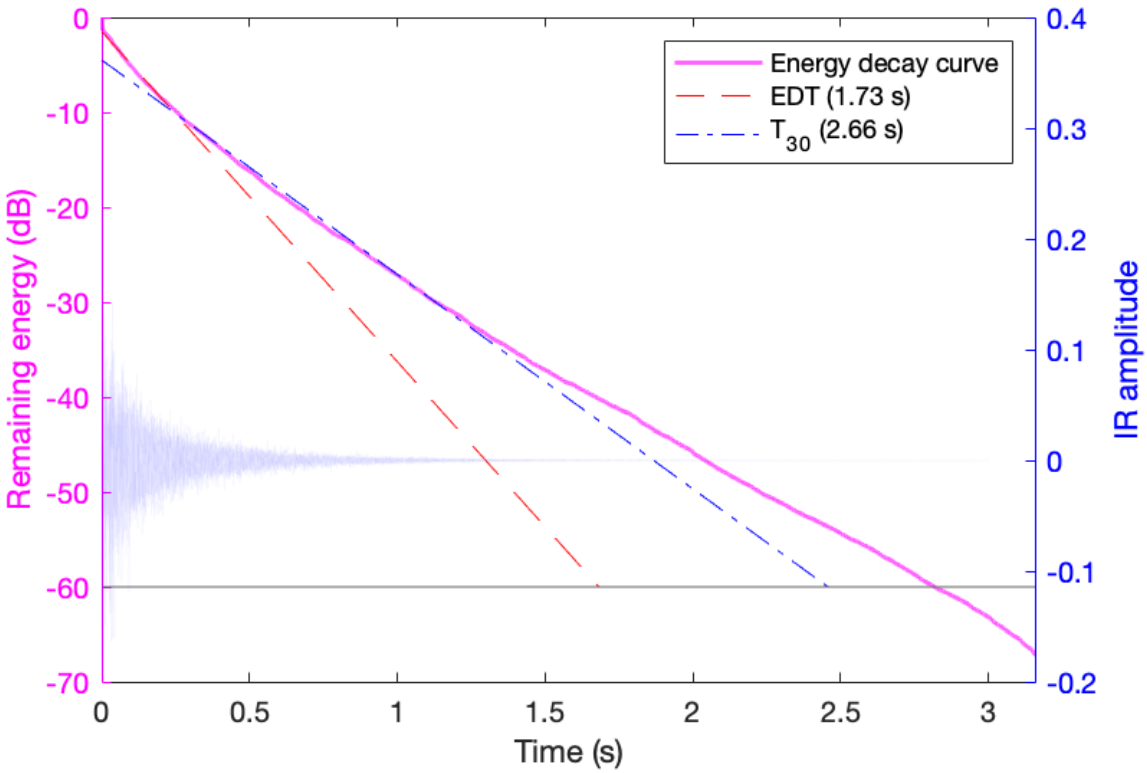


Figure 6 Energy decay curve example

Figure 6 shows examples of an impulse response, its energy decay curve E, and the lines of best fit used to calculate the *EDT* and T_{30}. Note that the *EDT* line closely tracks the decay curve only over the initial drop of 10 dB, while the T_{30} line tracks it closely only over the drop between -5 and -35 dB. The values of the two measures are shown visually by the location at which the line of best fit crosses the -60 dB threshold, shown here in grey. In this example, the *EDT* and T_{30} values diverge, indicating a non-linear decay. In this particular IR, sound energy decays more quickly in the initial portion relative to later portions. Typically, T_{30} and *EDT* measures are computed for eight versions of the IR filtered at octave bands.

2.2.2.1.2 Subjective level of sound (G)

$$G = 10 \log \frac{\int_0^\infty p^2\,(t)\,dt}{\int_0^\infty p_{10}^2\,(t)\,dt} \;\; dB$$

Equation 2 G

As is clear from Table 1 and Table 2, the subjective importance of an attribute related to the loudness or strength of sound in a hall is a robust finding in the perceptual literature. The feature proposed in the ISO model to predict this attribute, *G*, is essentially a ratio of the energy produced by the combination of a sound source and a room, to the energy produced by the same sound source without the room. As such, it measures very directly the effect of a particular acoustic enclosure on sound levels within it. If $p(t)$ is the impulse response of a room driven by a calibrated sound source, and $p_{10}(t)$ is the signal from the same sound source measured in a free field at a distance of 10 m, then G is defined as in Equation 2. As with the reverberance measures above, G is typically calculated for versions of an impulse response filtered at eight octave bands, leading to a distinct value of *G* at each octave band center frequency.

2.2.2.1.3 Perceived clarity of sound (C_{80})

$$C_{80} = 10 \log \frac{\int_0^{80} p^2\,(t)\,dt}{\int_{80}^\infty p^2\,(t)\,dt} \; dB$$

Equation 3 C_{80}

Much like G, the C_{80} measure associated with the perceived clarity of sound is also a ratio of two quantities. In this case, it measures the balance between early and late energy in the impulse response, with the temporal boundary separating the two regions

set at 80 ms. It is formally defined as in Equation 3, where the integration bounds are shown in the unit of milliseconds. The measure was initially proposed in Reichart et al. (1975).

2.2.2.1.4 ASW - apparent source width (J_{LF}, *IACC*)

Apparent source width is linked in the standard to two types of features: one calculated using a figure-of-eight microphone and the other calculated using a dummy head microphone.

The first measure, using the figure-of-eight, is the *early lateral energy fraction* (J_{LF}). This is the ratio of early energy arriving from the sides to the total early energy. As with the C_{80}, the boundary between early- and late-arriving energy is set at 80 ms. J_{LF} is defined as in Equation 4, where $p_L(t)$ is the impulse response captured by a figure-of-eight microphone with its null pointed toward the sound source. The measure was introduced by Barron and Marshall (1981).

$$J_{LF} = \frac{\int_5^{80} p_L^2(t)\ dt}{\int_0^{80} p^2\,(t)\ dt}$$

Equation 4 J_{LF}

The second measure, using a dummy head, is the *inter-aural cross correlation coefficient* or *IACC*. The *IACC* is the maximum value of the normalized cross correlation function between the left and right ear impulse responses, considered over lags between -1 ms and 1 ms. Again, only the first 80 ms of the binaural IRs are considered. Formally, if an interaural cross-correlation function is defined as in Equation 5, and p_l and p_r are the impulse responses captured at the left and right ears, then the *IACC* is the absolute value of the maximum of this function over lags specified by τ, as in Equation 6.

$$IACF(\tau) = \frac{\int_0^{80} p_l(t) p_r(t+\tau)\ dt}{\sqrt{\int_0^{80} p_l^2\ (t)\ dt \int_0^{80} p_r^2\ (t)\ dt}}$$

Equation 5 IACF

$$IACC = max|IACF|\ for - 1ms < \tau < 1ms$$

Equation 6 *IACC*

2.2.2.1.5 LEV - listener envelopment (L_J)

$$L_J = 10 \log \frac{\int_{80}^{\infty} p_L^2(t)\ dt}{\int_0^{\infty} p_{10}^2\ (t)\ dt}\ dB$$

Equation 7 L_J

Finally, listener envelopment is suggested by the standard to be correlated with the *late lateral sound level, L_J*. L_J measures the amount of sound energy arriving from lateral directions in the later portion of the impulse response, relative to the energy emitted by the sound source in a free field, as specified in Equation 7. The measure was introduced by Bradley and Soulodre (1995).

2.2.2.2 Signal features related to timbral attributes

Given that attributes related to timbre are present in virtually every investigation summarized in Table 1 and Table 2, the lack of such attributes in the ISO model is surprising, and indeed has been a source of criticism of the standard (J. S. Bradley, 2011). Despite their absence, however, a handful of studies over the years have posited their existence and have proposed objective signal features to predict them. While the perceptual structure of room reverberation timbre remains unclear, the timbral

attributes that have been proposed tend to cluster into three categories: those related to low-frequencies (e.g warmth, bassiness), those related to high-frequencies (e.g. brilliance, treble-to-mid frequency balance*)*, and those related to irregularities or unevenness in the frequency spectrum (e.g. timbral balance, colouration). These three broad attribute categories will be used to structure our discussion of signal features below.

2.2.2.2.1 Signal features related to low-frequencies (*BR, EBL)*

The earliest timbral predictor in the literature is perhaps Beranek's *bass ratio* (1962). This was proposed as the objective correlate of the subjective attribute of warmth. As shown in Equation 8, it is constructed from the ratio of low-frequency reverberation time (*T)* to mid-frequency reverberation time, where superscripts indicate octave frequency bands in Hz.

$$BR = \frac{T^{125} + T^{250}}{T^{500} + T^{1000}}$$

Equation 8 Bass ratio, BR

Later, as part of an extensive evaluation of objective predictors, Soulodre and Bradley collected subjective ratings on a low-frequency timbral attribute they described as "the strength of the bass or low-frequency sounds relative to mid-frequency sounds" (1995). They then examined this attribute's correlation with Beranek's bass ratio. The correlation between the two was quite low, leading them to propose a novel predictor dubbed the *early bass level (EBL)* to predict this attribute. *EBL* is defined as the sound strength *G* measured over the first 50 ms of an IR, averaged over the 125-500 Hz octave bands. This is expressed in Equation 9, where subscripts refer to time in milliseconds and superscripts refer to octave band center frequencies.

$$EBL = (G_{0-50}^{125} + G_{0-50}^{250} + G_{0-50}^{500})/3$$

Equation 9 Early bass level, EBL

2.2.2.2.2 Signal features related to high-frequencies (*TR, TR_{late})*

With respect to high frequencies, Beranek suggested that a ratio of high-frequency to mid-frequency *T* might correlate with his brilliance attribute (1962). This quantity might be called the *treble ratio* (*TR).* Hawkes and Douglas found some support for Beranek's hypothesis.

As with the bass ratio, however, Soulodre and Bradley also examined the predictive ability of a treble ratio in their 1995 article. They defined it as in Equation 10 and considered its correlation with a high-frequency attribute they described as "the strength of the treble or high-frequency sounds relative to mid-frequency sounds".

$$TR = \frac{T^{4\text{kHz}}}{(T^{1\text{kHz}} + T^{2\text{kHz}})/2}$$

Equation 10 Treble ratio, TR

In this case, their treble ratio predicted their attribute quite well. They found still better performance from a novel predictor, however, which consisted of a ratio of high- to mid-frequency energy in the late part of the IR, after 80 ms. This feature, which might be called the *late treble ratio (TR_{late})* is defined in Equation 11.

$$TR_{late} = 10\log\frac{\int_{80}^{\infty}(p^{4\,kHz})^2\,(t)\,dt}{\int_{80}^{\infty}(p^{1-2\,kHz})^2\,(t)\,dt}$$

Equation 11 Late treble ratio, TR_{late}

One open question regarding the features above, and reverberation timbre more generally, concerns the independence between attributes describing low-frequencies and attributes describing high-frequencies. In Table 1 and Table 2 it should be noted that most studies that reported a low-frequency attribute failed to report a high-frequency attribute, and those reporting a high-frequency attribute failed to report a

low-frequency attribute (e.g., Beranek's only timbral attribute was warmth; Hawkes and Douglas' only timbral attribute was brilliance). This raises the possibility that, within natural acoustic spaces, perceived levels of low frequencies and perceived levels of high frequencies might not vary independently. Rather, it may be the case that most variations in timbral balance between concert halls can be described on single axis with dark or bass-heavy at one pole and bright or treble-heavy at the other.

2.2.2.2.3 Signal features related to spectral irregularity (*DL)*

A third category of timbral attribute, which is discussed less frequently in the literature than those related to low or high frequencies, concerns the degree of perceived irregularity in the room's frequency spectrum. Weinzierl et al.'s colouration attribute would seem to belong to this category, as might Sabine's timbral balance attribute. One signal feature that attempts to predict this type of attribute is the *deviation of level (DL)* of Takahashi et al. (2008). *Deviation of level* is defined essentially as the standard deviation over frequency of an IR's smoothed magnitude spectrum. When this quantity is low, the spectrum is relatively flat, while a high quantity might indicate a jagged spectrum, as would be observed in the presence of comb filtering. As such, it seems plausible that such a feature might correlate well with an attribute such as Weinzierl's colouration. Calculation details of *deviation of level* are given in Takahashi et al.

2.2.2.3 Signal features related to diffusion

Although perceptual attributes related to acoustic diffusion are not nearly as prominent in the research reviewed in this chapter as attributes related to timbre, the perceptual consequences of diffusion in concert halls have nonetheless been a topic of interest to acousticians for some time, at least since the writings of Beranek in the 1960's. Barron, in 2005, summarized the state of knowledge on the perceptual effects of acoustic diffusion in this way:

> A mystery at present in concert hall acoustics concerns the subjective effects of substantial diffusing surfaces on the walls and ceiling of halls. There is some evidence that listeners prefer diffuse conditions […] but this is not conclusive. The state of diffusion remains to be satisfactorily quantified and no suggestions have been offered for how we perceive diffusion. (p. 163)

Although many aspects of diffusion perception remain enigmatic, in the decade-and-half since Barron's statement at least some attempts have been made to fill this lacuna. Studies on the perceptual consequences of acoustic surface diffusion have been conducted (e.g. Robinson, Pätynen, et al., 2013; Robinson, Walther, et al., 2013), and several signal features have been proposed to quantify the amount of diffusion captured in an impulse response. This section will focus on two such recently proposed features: the *number of peaks (NP),* and the *normalized echo density (NED) mixing time*.

Before proceeding, the term "diffusion" should be clarified, as it has a handful of distinct but interrelated meanings in the acoustics literature. On the one hand, a sound field is said to be "diffuse" or "directionally diffuse" when an equal amount of acoustical energy is arriving from all directions at the same time (Nolan et al., 2020). This quality of isotropic energy arrival is present in the late decay of reverberant spaces. It is this meaning that Beranek likely had in mind when he proposed diffusion as a subjective attribute of reverberation (see section 2.1.2). On the other hand, outside the context of isotropic sound fields, the diffusion-related terms are also used to describe a property of surfaces and reflections off these surfaces. Specifically, surfaces are said to be "diffusers" if they scatter reflected energy in time and space (Cox & D'Antonio, 2017). These scattered reflections are then said to be "diffuse" reflections.

2.2.2.3.1 Number of peaks (NP)

One signal feature aiming to quantify the amount of diffusion in an impulse response capitalizes on this temporal scattering of diffuse reflections. Specifically, the *number of peaks* feature measures the amount of diffusion by counting the number of local maxima in the rectified IR signal (Jeon et al., 2013). Intuitively, a room with only smooth, specular surfaces will give rise to reflections that are compact in time, and hence will generate few local maxima, while a room with diffusive surfaces will smear each reflection in time, resulting in many local maxima. Though conceptually simple, the number of peaks has proven to be quite effective in assessing changes in wall diffusivity (Bliefnick, 2016). The *number of peaks* is typically calculated separately in particular time-frequency regions, i.e. at octave band frequencies on both the early and late portions of an impulse response.

2.2.2.3.2 Normalized echo density (NED) mixing time

In addition to the directional- and surface-related meanings of diffusion, the term is also applied in a third context: to describe a particular kind of "late" reverberation, such as might be captured in the late tail of an impulse response (Gardner, 2002). So-called "diffuse reverberation" begins in an impulse response when the density of reflections is high enough that individual reflections can no longer be discerned. Late reverberation contrasts with early reverberation, which is characterized by temporal sparsity as low-order reflections arrive at the measurement position. Unlike early reverberation, late reverberation can be well approximated perceptually by random noise with a smooth power spectrum and an exponentially decaying temporal envelope (Jot et al., 1997).

The point at which reverberation transitions from "early" to "diffuse" is known as the *mixing time* (Blesser, 2001). Many signal features have been proposed to estimate this point, as reviewed by Lindau et al. (2012). One feature that is particularly effective for mixing time estimation is derived from an impulse response's so-called *normalized echo density (NED*). Computed over short temporal windows of an impulse response, the *normalized echo density* is defined as the number of samples outside of the window's standard deviation, normalized by the number which would be expected if the reverberation were diffuse (Abel & Huang, 2006). As such, the *NED* is a time-

varying feature that begins at a small value at the start of an impulse response and rises to a value of one when diffuse reverberation begins. The *NED mixing time* is the first point at which an IR's *NED* reaches one.

Relative to many other features discussed in this chapter, the perceptual correlates of mixing time remain poorly understood. Physically, a long mixing time indicates that the start of the impulse response contains widely spaced reflections. IRs with this characteristic are sometimes described as "rough" or "sputtery" (Abel & Huang, 2007). Conversely, short mixing times, which indicate closely spaced reflections and a more exponential decay, have been described as "lush" or "smooth" (Inglis, 2020). On the basis of these reports, it seems possible that an attribute such as Weinzierl's irregular decay might be partially explained by *NED mixing time*, though this relationship has yet to be confirmed.

2.2.2.4 Signal features from binaural auditory models (pRev, pClar, pASW, pLEV)

Until this point, the features discussed in this chapter have all shared the common property of being calculated from room impulse responses. Characterizing spaces in this way using IR-based features has been standard practice for at least a half-century. In this section an alternative approach to attribute value prediction will be presented that attempts to characterize sound fields using features derived not from a room's impulse response, but instead from a room's response to natural sounds sources, such as speech or music.

Basing signal features on a room's response to natural sound sources has the potential to increase attribute prediction accuracy, relative to IR-based features, because such features are able to account for interaction effects between source signals and rooms. A non-exhaustive list of such interaction effects includes the influence of sound source frequency on ASW (Mason et al., 2005) and the tendency of speech sources to produce stronger perceived reverberance than musical sources, even when convolved with the same IR (Teret et al., 2017). Such source-dependent percepts could potentially be predicted through a sophisticated analysis of the *wet* signal output from a convolution reverb effect, but would be impossible to predict through an analysis of the IR alone.

To analyze wet signals, rather than impulse response, the features in this section make use of a mathematical model of the human binaural hearing system: a so-called *binaural model*. Attempts to predict spatial attributes using binaural models date back several decades. Drawing on a central processing algorithm of Lindemann (1986), Becker proposed a binaural model to evaluate ASW (2002). Later, Rumsey et al. tuned a binaural model to evaluate spatial audio reproduction systems (2008).[10]

2.2.2.4.1 Schuitman

The work of Schuitman (2011) differs from the approaches above in that, unlike Becker, it provides a unified framework for predicting multiple reverberation attributes simultaneously, and, unlike Rumsey et al., it is designed specifically for the evaluation of natural sound fields, rather than reproduced ones. Building on a peripheral auditory model of Dau et al. (1996) and a central processing algorithm of Breebaart (2001), Schuitman uses biologically-inspired signal processing to produce estimates of the four attributes in the ISO 3382-1 model. Controlled experiments have shown that his model's predictions are often better aligned with perceptual ratings than the IR-based features suggested in the ISO standard (Lee et al., 2017). This is especially true for the spatial attributes of ASW and LEV (van Dorp Schuitman et al., 2013).

Schuitman's model follows the schematic shown in Figure 7. Its left- and right-ear peripheral processors consist of a bandpass filter to model resonance in the ear canal, a 41-channel gammatone filterbank to simulate the frequency resolution of the basilar membrane, a half-wave rectifier and low-pass filter to mimic transduction in the organ of Corti, and a frequency-dependent thresholding operation to model the absolute threshold of hearing. At the end of this signal chain, five feedback loops simulate the signal level adaption of neurons in the auditory pathway.

[10] This paragraph reviews efforts to predict ASW from wet signals using the outputs of peripheral auditory models. For completeness, it should be noted that attempts have also been made to predict ASW from wet signals *without* peripheral auditory modeling (e.g., Chernyak, 1968; Morimoto, 2002; Morimoto et al., 1993, 1994). While these papers represent important theoretical steppingstones in auditory science, none are directly applicable to the current work due to their specificity: the models they present tend to be valid only for one specific class of signal (e.g. noise or a particular musical motif) rather than for binaural stimuli in general.

The time-frequency data output from these peripheral processors is then input to a central processing algorithm. This algorithm determines a time-varying interaural time difference (ITD) in each frequency band, and sends this ITD information, along with the peripheral model outputs, to a central processor. The central processor performs a stream segregation, assigning time-frequency regions in the input signals to either an auditory foreground or auditory background stream. These foreground and background streams, each with its own frequency, level and ITD information, are then used to compute four objective features designed to predict the four attributes of the ISO 3382 model.

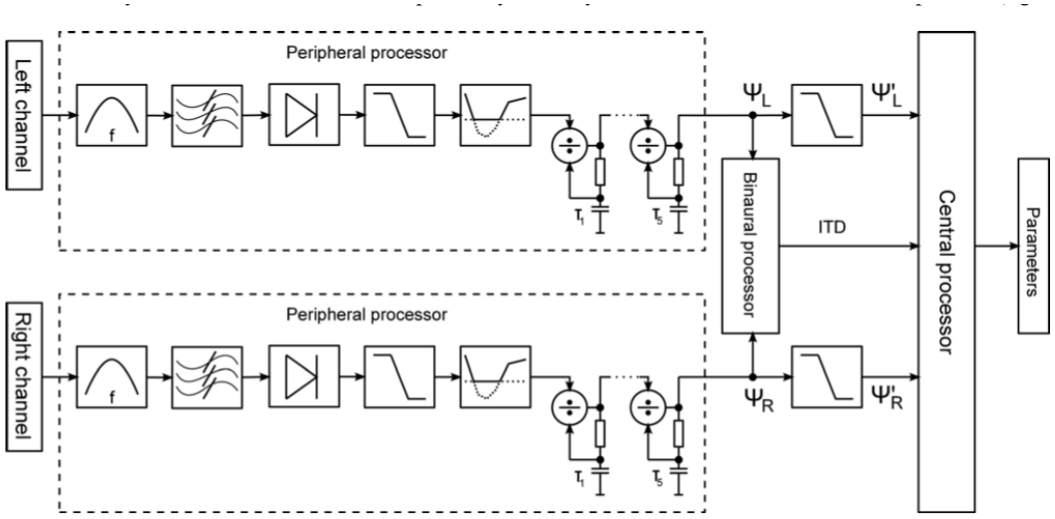


Figure 7 Schuitman's binaural model[11]

Schuitman's objective predictor for perceived reverberance, *pRev*, is simply the level of the auditory background stream. His predictor for perceived clarity, *pClar*, is the

[11] Reproduced from (van Dorp Schuitman et al., 2013), with the permission of the Acoustical Society of America.

ratio of foreground signal level to background signal level, a sort of perceptual direct-to-reverberant ratio.

His predictors for the spatial attributes of ASW and LEV draw on prior results that relate auditory spaciousness to fluctuations in ITD (Blauert & Lindemann, 1986). Specifically, his ASW predictor, *pASW*, is proportional to the standard deviation of ITD in the foreground stream, and his LEV predictor, *pLEV*, is proportional to the standard deviation of ITD in the background stream. Additionally, to model the effects of frequency on ASW, *pASW* also includes a term that depends on the level of low frequencies in the foreground stream. To account for the impact of level and reverberance on LEV (J. S. Bradley & Soulodre, 1995; Morimoto et al., 2007), *pLEV* also includes a term that depends on the level of the background stream.

2.3 Open questions relevant to perceptual search interfaces for IRs

This chapter examined research into perceptual models of room reverberation, with the aim of finding within the literature a model appropriate for organizing, in a convolution reverb context, a large collection of room impulse responses. The discussion showed that the model presented in ISO 3382-1 was a relatively faithful crystallization of the key findings in acoustics perception preceding it, and that it accounted relatively well for results produced after its publication. This model contains five attributes, one of which, subjective level of sound, is directly proportional to signal level. Since signal level is trivial to manipulate in an audio production context, a four-dimensional version of this model, with subjective level of sound removed, would seem to be a promising starting point for the IR search interface model.

2.3.1 Degree of inter-attribute and inter-feature correlation

A close reading of the literature, however, also reveals several open questions about this model. One question concerns the degree of independence between the model's attributes. Although the ISO attributes of reverberance, clarity, ASW and LEV are all phenomenologically distinct, section 2.1.2 noted that within real halls the attributes of

ASW and subjective level of sound tend to correlate. Similar strong relationships have also been observed between many objective features for reverberance and clarity (J. S. Bradley, 2011). Although correlated measures pose little problem in the context of venue assessment, they may be harmful in the context of search interface design. Here, correlated features may increase visual complexity without adding additional information to the display, thereby contributing to visual clutter in a way that may slow visual searches (Neider & Zelinsky, 2006). Understanding not only the salient attributes within an IR library but also the degree of correlation between these attributes would be useful in search interface design.

2.3.2 Reverberation timbre

A second question about the ISO model concerns attributes related to reverberation timbre. Section 2.2.2.2 noted that timbral attributes were nearly omnipresent in exploratory studies of reverberation perception, yet were entirely absent from the ISO model. This suggests that timbral attributes might vary within IR libraries, and that a model including them might be useful in IR search interface design. Indeed, when sound engineers evaluated a prototype IR search interface based on the ISO model, complaints about the absence of timbral features were frequent (Benson & Woszczyk, 2012).

Augmenting the ISO model with useful timbral attributes would require a detailed understanding of the perceptual structure of reverberation timbre, which is itself an open question. Three categories of timbral attribute seem to be present in the literature, as evidenced by an informal semantic clustering of attribute names. The first category contains attributes related to low frequencies, the second, attributes related to high frequencies, and the third, attributes related to spectral irregularity. One relevant and unanswered question about timbral attributes is the degree of correlation between these three attribute categories in natural spaces: is reverberation timbre truly a three-dimensional phenomenon or can it be well accounted for by fewer dimensions? Another question concerns the optimal objective signal features for predicting these attributes. Can the timbral features in the literature be improved upon?

2.3.3 Applicability of the ISO 3382-1 to IR search interfaces

A final question, so far undiscussed in this chapter but relevant to the current project, concerns the applicability of the ISO model to the task of reverb selection in mixing engineering. There are two reasons why the ISO model might be suboptimal in this context. For one, mixing tasks are carried out by trained sound engineers, and none of the studies above have specifically validated the ISO model with this population. As trained listeners, it is possible that sound engineers, through experience and enculturation, learn to attend to attributes of reverberation that are less audible to members of other professional communities.

One example of a sound engineering-specific attribute might be the type of perceptual variation associated with "diffusion" controls on algorithmic reverb units. These controls are said to manipulate an attribute described as "grainy" or "metallic" at one pole and "lush" or "smooth" at the other (Inglis, 2020). This attribute is not present in the ISO model, and it seems likely that synthetic reverb would exhibit greater variation in this attribute than natural reverb. Nonetheless, it is plausible that experienced sound engineers, having explored the attribute in algorithmic reverb, might become more sensitive to it than populations of acousticians or concertgoers, and might become capable of attending to subtle variations of it in natural IR libraries.

Another reason that the ISO model might be an imperfect abstraction for the task of impulse response library searching concerns loudness. Whereas subjective level of sound varies considerably between concert halls and is an important source of perceptual differences between them, sound level is trivially easy to control at a mixing desk. One might expect that reverb effects previewed by engineers during mixing sessions would be relatively well matched in loudness. This loudness matching might allow more subtle attributes, not present in the ISO model, to become more salient. If this were to occur, the ISO model might prove an insufficiently detailed description of the perceptual variations in convolution reverb outputs.

2.3.4 Summary

In summary, then, while the ISO 3382 model would appear to be successful in characterizing concert hall-like reverberation for a broad population of listeners, it may not be the optimal model for describing the perception of reverberation by sound engineers carrying out mixing tasks. There are three reasons for this: first, it fails to characterize timbral attributes; second, it may fail to characterize attributes to which sound engineers are uniquely sensitive; third, it may fail to characterize attributes which only become salient once loudness differences between stimuli are removed. The discussion above suggests that a focused investigation of reverberation perception by sound engineers, and, in particular, of loudness-matched reverberation, might be fruitful for informing the design of a perceptual model for an IR search interface. If the perception of loudness-matched reverberation by mixing engineers were better understood, this knowledge could be applied to adapting or refining the ISO 3382 model to better serve the specific needs of this population.

Finally, with respect to objective signal features, research conducted since the publication of the 2009 ISO standard suggests that features derived from binaural models may be more effective than IR-based features for predicting some ISO model attributes. This is especially true for the spatial attributes of ASW and LEV.

3 EXPLORATORY ANALYSIS OF REVERBERATION ATTRIBUTES

The previous chapter reviewed research into perceptual models of natural reverberation. It concluded that the model presented in ISO 3382 was a relatively faithful summation of research results. It also raised the possibility, however, that the ISO model might fail to include some reverberation attributes relevant to mixing engineering tasks. Two speculative reasons for the absence of these attributes were proposed. First, it was possible that there existed certain attributes of reverberation to which professional sound engineers were more sensitive than the general population. The studies informing the ISO model did not generally investigate perceptual idiosyncrasies of the sound engineering community, so it is unsurprising that these attributes might be excluded. Second, it was possible that there existed subtle attributes of reverberation that only became apparent when loudness differences between sound fields were removed. The chapter concluded by suggesting that an exploratory investigation into the perception of loudness-equalized reverberation by sound engineers might be fruitful. Such a study might expose any such mixing engineering-specific attributes. Identifying and understanding such attributes would have obvious value in informing IR search interface design.

This chapter describes such an exploratory investigation. Through a two-part experiment involving loudness-matched reverberation, it seeks to identify salient sources of perceptual differences in natural reverberation beyond those in the ISO model. To do so, it first identifies set of putative reverberation attributes through an objective analysis of algorithmic reverb presets. These attributes are said to be "putative" because, although they appear to correspond to identifiable signatures in the space of objective signal features, it is unclear whether these feature signatures have

perceptual correlates. The perceptual validity of these putative attributes will then be assessed in a subjective listening test.

The investigation will proceed in two parts.

In the first part (3.1), a collection of output signals from algorithmic reverb units will be analyzed. These output signals, each corresponding to an algorithmic reverb *preset*[12], will first be classified linguistically. The preset names will be reduced to a small set of descriptive labels, such as “bright”, “dark” and “vocal”. A statistical algorithm will then be used to build a predictive model that relates each label to a set of signal feature values.

In the second part of the chapter (3.2) these predictive label models will be evaluated in a listening test. A double-blind experiment will determine whether any of the models, trained on the collection of algorithmic presets, are also able to reliably predict perceptual variations within a library of natural IRs.[13] For those models that do predict perceptual variations in the natural library, subjects will describe the nature of the variation in words. An examination of the subjects’ verbal descriptions will then shed light on the dimensions of perceptual variation within the natural IR library.

The final part of the chapter (3.3) will review what the investigation suggests about the perceptual structure of reverberation in loudness-equalized contexts, in particular with respect to attributes beyond the four in the ISO 3382 model.

3.1 Predictive modelling of algorithmic reverb preset labels

The first part of the chapter will focus on building predictive models of words used to describe algorithmic reverb presets. The investigation will make use of a large collection of such presets. The method of assembly of the collection will first be

[12] Presets are discussed in section 1.1.1.

[13] This thesis uses the term "natural" to refer to IRs measured in physical acoustic spaces such as concert halls. Natural IRs contrast with "synthetic" IRs, which are the measured at the outputs of reverb algorithms. See discussion at the start of Chapter 2.

explained, followed by the method for associating descriptive words (i.e. labels) with items in the collection. The chapter will follow with a brief discussion of signal features and statistical techniques, and will then present predictive models for 12 preset labels.

3.1.1 Assembling a library of algorithmic reverb presets

To conduct the investigation, a library of 924 algorithmic reverb presets was assembled. The library included presets from 12 reverb units released between 1982 and 2008. The presets were stored as stereo impulse responses.

The presets were collected from several websites. Those for the *AMS RMX 16* were found on a blog[14] while those for the *Bricasti M7* were made freely available by a company specializing in the distribution of reverb impulse responses.[15] The rest were purchased for a small fee from a second company.[16] These three sites were identified from postings on a sound engineering discussion forum.[17]

No particular criteria were used to determine the composition of the library other than the ease-of-availability of the IRs. All preset IRs that could be found on the internet, following a modest amount of web searching, were included. No effort was made to restrict the set of devices to those that were especially popular, or had received critical acclaim. Rather, the mere fact that some individual had taken the time to capture the device's IRs was interpreted as a weak signal that the device was of interest to the sound engineering community, and that the words used to describe the presets would be broadly consistent with the understanding of these words in the minds of sound engineers.

[14] http://fakekrates.tumblr.com/post/102613258467/ams-rmx-16-impulse-responses
[15] http://www.samplicity.com/
[16] http://signaltonoize.com/
[17] https://www.gearspace.com/

3.1.1.1 Preset "naturalness" screening

To focus the investigation on perceptual attributes that were likely to vary in natural reverb, the 924 IRs were screened for decay time and decay linearity. It was expected that many of the presets, having been generated by reverb algorithms, would have implausibly long decay times or non-linear decay slopes. Accordingly, presets with T_{30} values above 4.7 seconds or a *degree of non-linearity* above 2.1% were removed from the library.[18] Following this screening, 702 presets remained.

These 702 presets, along with their associated reverb units and years of release, are shown in Table 3.

[18] The *degree of non-linearity* is an objective measure of decay slope linearity defined in ISO standard 3382-2 (International Organization for Standardization, 2008). It is inversely related to the correlation coefficient between the observed energy decay curve (Equation 1) and the best-fitting linear decay line. IRs with perfectly linear decays have a degree of non-linearity of 0%.

Table 3 Devices represented in the preset library

Release date	Manufacturer	Device	Num. presets
1982	AMS	RMX16	8
1985	Lexicon	PCM 70	10
1986	Lexicon	480L	21
1987	Quantec	QRS-XL	2
1989	Lexicon	300	14
1992	Eventide	H3000	53
1993	Yamaha	SPX 990	21
1995	TC	M5000	13
1996	Lexicon	PCM 90	82
1999	TC	6000	51
2006	Lexicon	MX200	43
2008	Bricasti	M7	89

No formal efforts were made to verify the authenticity and provenance of each preset IR; it was assumed that each was a faithful recording of the associated device's output. Many were informally auditioned, however, and no obvious capture errors or defects were heard.

3.1.1.2 Identification of popular labels

Each preset in the library had both an audio component and a text component. The preset's audio component consisted of a stereo impulse response; its text component consisted of its name, as recorded in the IR file's name and relative path.

From each of the 702 presets' names, a set of zero or more "labels" was extracted. Labels were defined as descriptive words that met certain criteria. Specifically, to be considered a label, a word needed to be either: an adjective describing the sound of the preset (e.g. "bright", "warm"), a noun indicating a type of space ("hall", "chamber"), a noun indicating a reverb generation method ("plate", "spring"), or a noun indicating a type of audio material well suited to the preset ("vocal", "drums"). Labels were extracted manually from the presets' names by the author.

Additionally, label spellings were normalized, and abbreviations were expanded. Preset names frequently contained abbreviated terms, such as "A.Gtr" for "Acoustic Guitar" or "Amb." for "Ambience", presumably due to space constraints on the displays of the original rack mounted devices. These abbreviations were expanded to their more standard forms. Nonstandard spellings were also normalized (e.g. "brite" to "bright").

Once labels were determined for the presets in the library, the number of occurrences of each label was counted. Figure 8 shows the set of labels that each occurred more than ten times in the collection. Occurrences are color-coded by reverb unit manufacturer.

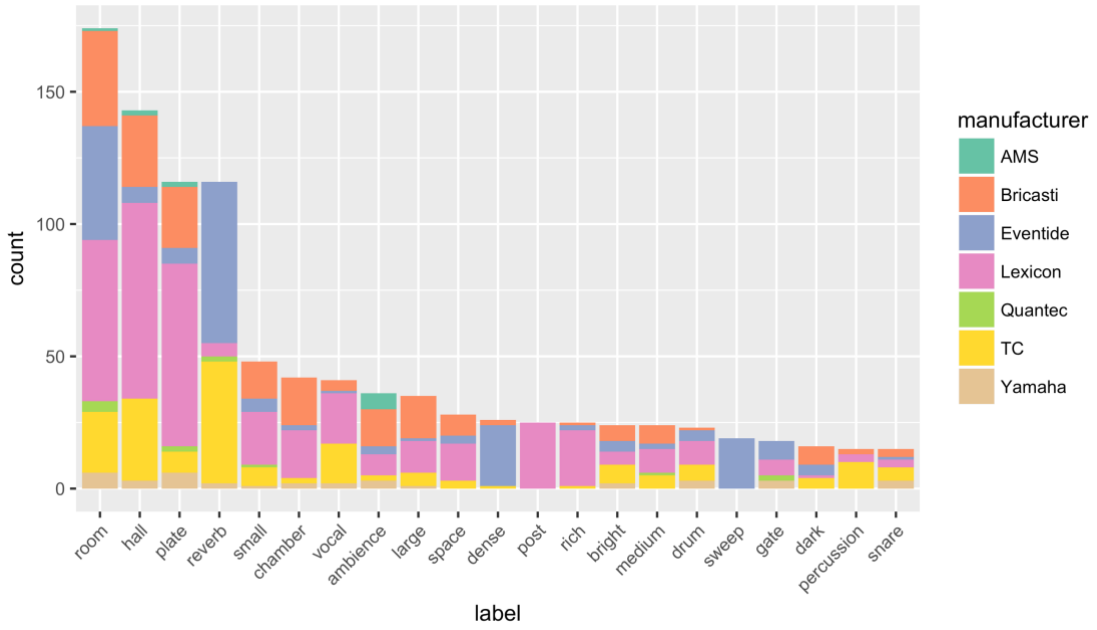


Figure 8 Preset labels with more than 10 occurrences in the library

3.1.1.3 Choosing a set of labels to analyze

Having determined which labels occurred most frequently in the collection, the next task was to reduce this set of 21 labels down to a more manageable number. The final set of labels to investigate in detail was chosen subjectively, with the aim of including those that seemed most likely to correspond to attributes not present in ISO 3382-1. This section outlines the reasoning used to trim down the initial set of 21 labels.

The goal of investigating perceptual attributes of natural reverb led to excluding the label "gate". It was expected that most natural IRs would have smooth decays, rather than non-linear, gated decays. As such, gatedness was not an attribute that was expected

to vary in a natural IR collection. Likewise, the label "sweep" was excluded since the defining perceptual characteristic of algorithmic “swept” reverbs – a modulated delay time – was unlikely to be present in natural reverb (as well as being impossible to capture in a time-invariant impulse response).

The goal of investigating novel attributes – that is, attributes beyond those in ISO 3382 – also led to excluding labels that seemed obviously related to perceived reverberance. These labels included "room", "hall", "small", "large" and "medium". Intuition, along with preliminary investigations, suggested that these labels would be strongly associated with particular ranges of reverberance-related features such the T_{30} and EDT. Investigating the perceptual characteristics of these labels would only have reaffirmed the centrality of perceived reverberance to room acoustics perception, an unsurprising and uninteresting result.

Also excluded were the labels "reverb" and "post", which seemed unspecific and unlikely to correspond to a single perceptual attribute. "Post" is generally used to denote IRs intended for audio post-production, such as reverberation from car interiors, bathrooms and offices. The wide variety of spaces collected under this label seemed unlikely to share a common perceptual attribute which could be captured by our simple model building algorithm.

Removing the nine labels discussed above left a set of twelve labels appropriate for further study. The twelve labels selected for the investigation were "plate", "chamber", "vocal", "ambience", "space", "dense", "rich", "bright", "drum", "dark", "percussion" and "snare".

3.1.2 Building predictive models

This section explains the methods used to develop predictive models of the preset labels. These models take objective signal features as input, and yield label probabilities as output. The signal features used in modeling will be reviewed first, followed by the statistical techniques used to the create the models. Each resulting label model will then be presented in detail.

3.1.2.1 Signal features

The signal features calculated for each preset fell into two broad categories. In the first category were features calculated on the 100% wet output from the preset, that is, the result of the convolution between the preset's IR and a source signal. These were referred to as *wet* features. In the second category were features measured directly on the preset's stereo impulse response. These are referred to as *IR* features.

A complete list of features appears in Table 4. Most were described in detail in the previous chapter; others, as indicated, are described in Appendix B. Unless specified as wet, all features were calculated on the IR signal alone. Wet features, by contrast, were created by convolving the preset IR with one of three sound sources (orchestra, chorus, jazz voice) and analyzing the output signals.[19] The three sound sources are detailed in appendix C.

3.1.2.1.1 Clustering by octave band

Many of the IR features were initially calculated in eight versions, one for each of eight octave bands with center frequencies between 63 and 8000 Hz. To reduce redundancies between the features, the set was subjected to a cluster analysis. Features from octave bands that clustered together were grouped into composite features. In the case of *EDT*, for example, values for the 63 through 1000 Hz bands were relatively highly correlated, and so were averaged and combined into a single composite feature called *EDT 63-1k*. Other composite features are indicated in Table 4.

[19] The sound source used to calculate wet features was varied according to the label being investigated. Features for the "vocal" label were measured using the jazz voice source; those for the "dark" and "bright" labels were measured using the orchestra source. All other labels used the chorus sound source. Orchestra was used for "dark" and "bright" to keep the source material broadly consistent with prior research into timbral attributes, much of which also used orchestral material (e.g., Beranek, 1962; Soulodre & Bradley, 1995). The chorus source was chosen for the other labels in the hopes that its wide bandwidth and complex spectro-temporal variation would help to highlight differences between presets in the space of signal features.

In the case of the *number of peaks* feature, values for each octave band were standardized (i.e., converted to *z*-scores) before averaging. This was done to keep the average from being dominated by higher frequency bands, where the number of peaks generally took on values that were orders of magnitude higher than in lower frequency bands.

3.1.2.1.2 Mono features vs binaural features

Both the preset IRs and the wet signals were composed of two channels (left and right). Features which were designed to operate on two-channel signals (i.e. those in the *IACC* and *Binaural Model (Wet)* groups) used these two channels as input.[20] All other features were designed to operate on single-channel signals. For these monophonic features, the left and right channels were added together, and this summed signal was used to calculate the feature.[21]

[20] Note that these "two-channel" features were all originally conceived to operate on binaural signals, such as might be captured by a dummy head microphone, rather than on synthetic reverberation intended for loudspeaker reproduction. The rationale for applying these features outside of their intended context is given in Appendix B.

[21] This summation was performed to ensure that information from both channels was present in the features. In retrospect, however, this approach may have been misguided, as time domain summation may have introduced comb filtering or other spectral artifacts that would have been imperceptible in headphone listening. A better method for including information from both channels in the features might have been to calculate features on both channels separately and then average the two values. Alternatively, such artifacts could also have been avoided by choosing to analyze only a single channel (left or right) from each stereo pair.

Table 4 Signal features used in modeling

Group	*Abbreviation*	*Description*
EDT	EDT 63-1k	Early Decay Time, average of 63 to 1000 Hz bands
	EDT 2-4k	…, average of 2000 and 4000 Hz bands
	EDT 8k	…, 8000 Hz band
T30	T30 63-500	T30, average of 63 to 500 Hz bands
	T30 1-2k	…, average of 1000 and 2000 Hz bands
	T30 4-8k	…, average of 4000 and 8000 Hz bands
C80	C80 125	C80, 125 Hz band
	C80 250-2k	C80, average of 250 to 2000 Hz bands
	C80 4k	
	C80 8k	
IACC	IACC	Average of 1-2k bands, as in Beranek and Okano
	IACC late	..., post 80 ms (see appendix B)
Timbral (IR)	BR (EDT)	Bass Ratio, calculated on the same temporal region as the EDT
	BR (T30)	..., calculated on the same temporal region as the T30
	TR (EDT)	Treble Ratio, calculated on the same temporal region as the EDT
	TR (T30)	Treble Ratio, calculated on the same temporal region as the T30
	EBL	Early Bass Level
	LTR	Late Treble Ratio
Diffusion-related	NP early 63-125	Number of Peaks (early), standardized average of 63 and 125 Hz bands
	NP early 250-500	…, standardized average of 250 and 500 Hz bands
	NP early 2k-4k	…, standardized average of 2000 and 4000 Hz bands
	NP late 63-125	Number of Peaks (late), standardized average of 63 and 125 Hz bands
	NP late 250-500	…, standardized average of 250 and 500 Hz bands
	NP late 1-4k	…, standardized average of 1000, 2000 and 4000 Hz bands
	NP late 8k	…
	NED Mixing Time	Normalized Echo Density mixing time
Attack Time	Log Attack Time	Log Attack Time (see appendix B)
Non-linearity in decay	Degree of Curvature	ISO 3382-2 (see appendix B)
Binaural model (Wet)	pRev	Reverberance measure
	pClar	Clarity measure
	pASW	ASW measure
	pLEV	LEV measure
Timbral (Wet)	Spectral Slope	Calculated with the Timbre Toolbox (see appendix B)
	Spectral Skew	Timbre Toolbox (see appendix B)
	Spectral Decrease	Timbre Toolbox (see appendix B)
	Spectral Flatness	Timbre Toolbox (see appendix B)

3.1.2.2 Statistical methods

To build predictive models for the twelve selected labels, two statistical techniques were employed: logistic regression (LR) and supervised principal components analysis (SPCA) (Bair et al., 2006). A brief introduction to each follows.

3.1.2.2.1 Logistic regression

In a logistic regression model, the probability of a binary outcome, such as a label being present on an IR, is modeled as a function of an input variable using a logistic curve. Logistic curves increase monotonically from zero to one and have a vaguely "S"-like shape. As an example, Figure 9 shows a logistic regression of the *spectral decrease* feature on the "dark" label, for 40 of the presets. As the *spectral decrease* grows, the modelled probability of the "dark" label, $p(dark)$, rises. The logistic curve relates the feature value to the label probability.

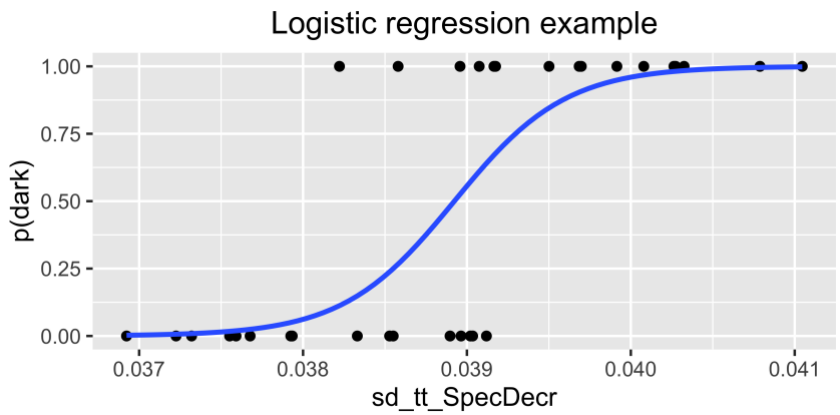


Figure 9 Example of a logistic regression model

Mathematically, a logistic regression model has the form of Equation 12, where x is the input feature, $f(x)$ is the probability of the label, and $\hat{\beta}_0$ and $\hat{\beta}_1$ are model parameters estimated from the data. In the example above, $\hat{\beta}_0$ is -114.7 and $\hat{\beta}_1$ is 2947.[22]

$$f(x) = \frac{1}{1 + e^{-(\hat{\beta}_0 + \hat{\beta}_1 x)}}$$

Equation 12 Logistic Regression

3.1.2.2.2 Supervised Principal Component Analysis (SPCA)

The models discussed in this chapter are similar to the logistic regression model presented above, except that rather than using a single feature as input, such as the spectral decrease, the models use a combination of several features. Mathematically, this means that the input variable, x, is replaced by a weighted combination of features, t. This is shown in Equation 13. Here $x_1, x_2, \ldots, x_p$ are the features used in the model, and $w_1, w_2, \ldots, w_p$ are the weights applied to these features.

$$t = w_1 x_1 + w_2 x_2 + \cdots + w_p x_p$$

Equation 13 Feature weights

Aside from this variable substitution of x for t, a Supervised Principal Components Analysis-Logistic Regression model (SPCA-LR) is identical to the univariate logistic regression shown in Equation 12. SPCA-LR is different from simple logistic regression only in its use of the weighted feature combination t in place of single feature x. The general form of an SPCA-LR model is shown in Equation 14.

[22] These model parameters are larger than those typically seen in practice because, in this example, the input variable x has not been standardized. Its magnitude is very small, taking on values between about 0.037 and 0.041, and the model parameters $\hat{\beta}_0$ and $\hat{\beta}_1$ must be large to compensate.

$$f(x) = \frac{1}{1 + e^{-(\hat{\beta}_0 + \hat{\beta}_1 t)}}$$

Equation 14 SPCA Logistic Regression

One advantage of the SPCA-LR approach, relative to a univariate approach, is that it allows a model to show relationships between a label and multiple features. If other features besides *spectral decrease*, for example, were also useful for predicting "dark", these other features might also appear in the model. The exact procedure by which features and weights are chosen for each label model is detailed in Appendix D.[23]

3.1.3 Label modelling results

The label and feature data presented earlier was submitted to the SPCA-LR algorithm, generating models for each of the twelve labels. Randomization tests were then conducted on the resulting models to check whether they had succeeded in capturing meaningful relationships between the features and labels, and were able to predict the labels at rates better than chance (Golland & Fischl, 2003). The randomization tests, described in Appendix D, produced a *p*-value for each model. These *p*-values are shown in Table 5.

[23] An alternative method for investigating relationships between a label and multiple features is multiple logistic regression. Despite being a more standard technique, multiple logistic regression was difficult to apply here due to the data's high dimensionality. The rationale for using SPCA-LR instead of multiple logistic regression is explored in Appendix D.

Table 5 Model *p*-values

	p
plate	0.00
dense	0.00
ambience	0.00
dark	0.00
chamber	0.00
bright	0.01
vocal	0.01
rich	0.01
drum	0.04
snare	0.57
space	0.62
percussion	0.86

Using a significance threshold of $\alpha = 0.05$, the permutation tests found that nine out of the twelve models had succeeded in capturing a relationship between features and the label. For the three models with p values greater than 0.05 (*space*, *percussion*, *snare*), the model-building procedure was not able to find any relationship between the labels and features. The remainder of the chapter will focus on the former cases; the three unsuccessful models will not be discussed further.

The precise structure of the nine successful models is shown in Figure 10. For each model, the top panel shows the regression coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$, while the bottom panel shows the weights $(w_1, w_2, \ldots, w_p)$ applied to each feature. A blank cell indicates that a feature was not selected for a given model by the SPCA-LR algorithm.

	vocal	dark	plate	bright	dense	rich	chamber	ambience	drum
$\hat{\beta}_0$	-2.96	-4.38	-1.79	-3.61	-3.81	-4.05	-2.97	-3.03	-3.62
$\hat{\beta}_1$	0.58	0.73	0.82	0.40	0.71	1.63	0.66	0.32	0.53
	vocal	**dark**	**plate**	**bright**	**dense**	**rich**	**chamber**	**ambience**	**drum**
EDT 63-1k									
EDT 2-4k								-0.39	
EDT 8k	0.44								
T30 63-500									
T30 1-2k									-0.68
T30 4-8k				0.04					-0.67
C80 125									
C80 250-2k									
C80 4k	-0.48								
C80 8k	-0.51								
IACC early					0.71	-0.71			
IACC late						-0.71		0.64	0.29
BR (EDT)				-0.07					
BR (T30)				-0.19					
TR (EDT)		-0.48	0.25						
TR (T30)			0.29	0.25					
EBL (SB)	-0.71			-0.19					
TR (SB)	0.41	-0.63	0.43	0.42	-0.17		-0.71		
NP early 63-125				-0.16					
NP early 250-500									
NP early 2k-4k									
NP late 63-125				-0.23					
NP late 250-500							0.71		
NP late 1-4k									
NP late 8k					-0.39				
NED Mixing Time							-0.57		
Log Attack Time				-0.07					
Curvature									
pRev									
pClar									
pASW					-0.71				
pLEV								-0.67	
Spectral Slope			0.49	0.47					
Spectral Skew			-0.47	-0.39					
Spectral Decrease	-0.71	0.60	-0.46	-0.44					
Spectral Flatness				0.35					

Figure 10 Label models

One notable aspect of Figure 10 is the structural similarity of many of the models. Some groups of models, such as bright and plate, appear to be composed of similar features. This suggests that the predictions of some models might be correlated.

3.1.3.1 Model output correlation

Model correlations were explored more fully in Figure 11. In this tree diagram, each model is represented by a "leaf" at the bottom of the image. Leaves are joined together by branches; the height of each branch shows the amount of correlation between model outputs.[24] Most of the models (all except *chamber*) cluster into three groups. *Vocal*, *dark*, *plate* and *bright* form group 1; *dense* and *rich* form group 2; *ambience* and *drum* form group 3. These similarities in the numerical outputs of the models suggest that there may be perceptual similarities in their outputs as well. These perceptual similarities will be considered in the perceptual evaluation detailed in the next section.

[24] More precisely, the branch height shows the absolute value of the Spearman rank correlation, subtracted from one. To create the tree, 31564 wet signals were created by convolving the four source signals with 7891 IRs from the Spacebuilder Library. These signals were then run through the predictive models, generating a probability for each of the nine labels. The predictions were assembled into a matrix with nine columns and 31564 rows. Spearman's rank correlation was calculated for each pair of columns, and the resulting correlations were converted to distances (i.e., subtracted from one). The distances were then subjected to an agglomerative hierarchical clustering analysis with complete linkage.

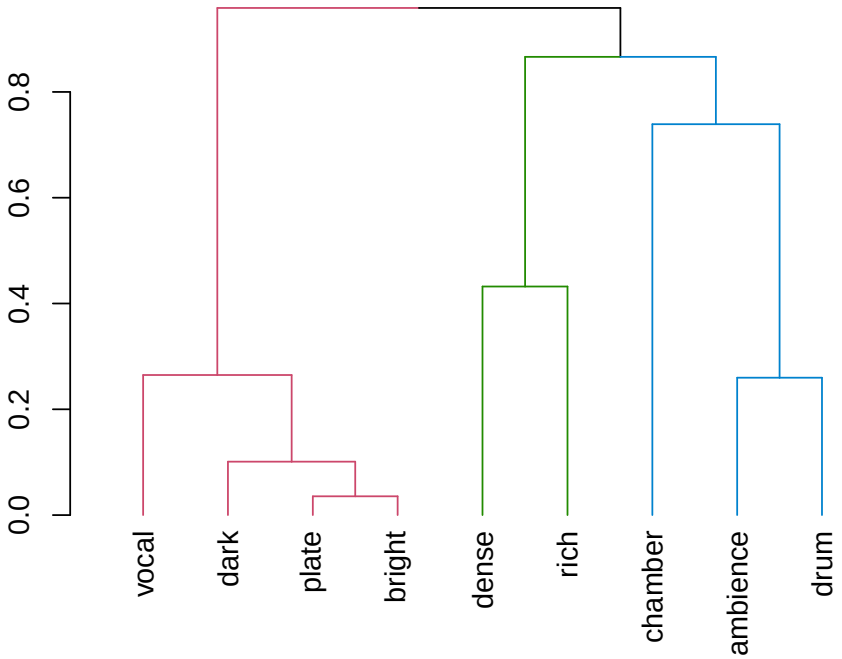


Figure 11 Dendrogram of model correlations

3.2 Perceptual evaluation of label models

In the previous section, predictive models for nine popular algorithmic reverb preset labels were developed. These labels were "vocal", "dark", "plate", "bright", "dense", "rich", "ambience" and "drum". The fact that some presets carried these labels and others did not, along with the fact that each label exhibited a particular acoustic signature, suggests that the labels may correspond to dimensions of perceptual variation

within the collection of presets. That is, it is plausible that the labels convey genuine information about the perceptual qualities of presets: that drum-labelled presets share certain qualities that distinguish them from non-drum presets, and likewise for the other eight labels. These plausible dimensions of perceptual variation - *vocalness, darkness, plateness, brightness, density, richness, ambience-ness,* and *drumness* - might be called "putative attributes" of reverberation.

In this section these putative attributes, and the label models associated with them, will be used to investigate the perceptual structure of natural reverberation. Specifically, the models will be used to attempt to identify salient attributes of reverberation that are not present in ISO 3382-1. For each of the nine label models, natural IRs will be found with relatively low and high probabilities of the label, as predicted by the model. That is, natural IRs will be found with both low predicted probabilities of the "drum" label and high predicted probabilities of the "drum" label. A listening test will then determine whether label probability differences translate into perceptual differences: whether natural IRs with a low probability of "drum" do indeed "sound different" from natural IRs with a high probability of "drum".

A listening test was constructed to answer three specific questions about the label models.

First, the test sought to determine which of the nine models were perceptually discriminable. For which models did signals with a high probability of a label sound different from signals with a low probability of a label? For example, were signals with a high probability of "dark" perceptually different from signals with a low probability of "dark"?

Second, for those models that were discriminable, what sorts of words did recording engineers use to describe differences in probabilities? How did engineers verbally characterize the difference between signals that were more and less likely to have to have the "drum" label? Was this difference described as a literal variation in "drumness", or were other words used?

Finally, of the nine models, did any describe perceptual attributes that weren’t present in ISO 3382-1? Were vocalness, darkness, plateness, brightness, density, richness,

ambience-ness and drumness similar to perceived reverberance, perceived clarity, ASW and LEV, or did any seem distinct?

The next section describes the listening test's methodology. The section afterward (3.2.2) presents the test's results and discusses the first two questions listed above. The final question, about the implications of the findings on the perceptual attributes of natural reverberation, will be addressed at the end of the chapter, in section 3.3.

3.2.1 Methodology

This section describes the listening test used to investigate the label models. The paragraphs below outline its subjects, stimuli, user interface and structure.

3.2.1.1 Subjects

Fifteen subjects were recruited from the Sound Recording Area of McGill University's Schulich School of Music. All subjects were graduate students or faculty members at the School, and all reported at least two years of audio engineering experience. Distributions of subject age, academic level and audio engineering experience are shown in Figure 12. The labels "SR1" and "SR2" refer to the first and second year of a two-year master's degree.

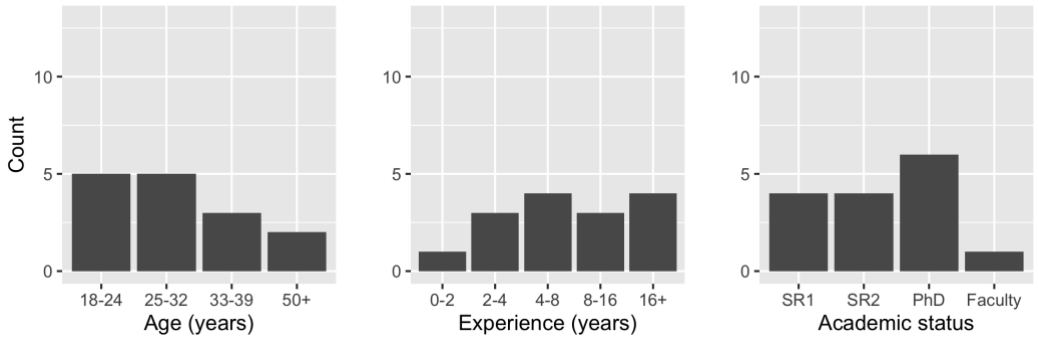


Figure 12 Listening test subject demographics

3.2.1.2 Stimuli

All stimuli used in the experiment consisted of a "100% wet" reverb signal, i.e. the output of a convolution between a stereo IR and a monophonic sound source. The IRs in the experiment were drawn from a subset of the Spacebuilder library (see Appendix A). The sections below describe sound sources used, as well as the method used to select stimuli and arrange them into groups for experimental trials.

3.2.1.2.1 Sound Sources

Four monophonic sound sources were used to generate stimuli. The four were intended to represent a broad range of musical material while still keeping the experiment to a reasonable duration. The stimuli, detailed in Appendix C, are referred to here as chorus, drums, jazz voice and orchestra.

3.2.1.2.2 Stimulus grouping

The listening test employed a triadic comparison paradigm, meaning that stimuli were presented to subjects in groups of three. The IRs used to generate stimuli were selected from the Spacebuilder library using a custom sampling algorithm. The algorithm, detailed in Appendix E, attempted to sample randomly from the library while also

ensuring that the three stimuli in each trial exhibited maximal variation in predicted label probabilities but minimal variation in ISO model attributes. The rationale for minimizing variations in the ISO attributes was to focus the subjects' attention on the putative attributes under investigation: vocalness, darkness, plateness, brightness, density, richness, ambience-ness and drumness.[25]

All stimuli were loudness equalized to -18 LUFS (International Telecommunications Union, 2011).

3.2.1.3 Experimental interface

Subjects completed the experiment using a software interface, a screenshot of which is shown in Figure 13. The interface allowed subjects to compare groups of three stimuli. Audio playback could be started and stopped using the spacebar, and the arrow keys on the keyboard allowed for seamless switching between stimuli. Stimuli were auditioned over headphones. Subjects were able to control playback volume and were instructed to set it to a comfortable level.

Subjects were instructed to listen to all stimuli in each trio, and then indicate the one that was most different. This chosen stimulus would presumably exhibit some quality that was absent from the other two stimuli. Subjects were then asked to describe this distinctive quality using a word or a short phrase. As an example, subjects were invited to consider a hypothetical trial in which two stimuli were highly reverberant and one was dry. In this case they were instructed to select the dry stimulus, and to write the word "dry" in the text box.

Subjects were invited to make similarity judgements on the basis of any perceptual attribute with the exception of one: the angular position of the sound source image. This restriction was chosen due to the relatively large amount of variation in apparent sound

[25] This discussion assumes that the putative attributes were all unidimensional. In the event that any of the putative attributes were multidimensional, the goal of minimizing variations in the ISO attributes was to focus subjects' attention on any components of the putative attribute that could not be explained by the ISO model.

source angle within the library. Sound source angle is an extremely salient attribute, and without accounting for it, the researchers worried that it might dominate similarity judgements. That is, that subjects would consistently attend to and report only on sound source angle, rather than other more subtle attributes. These variations in sound source angle arose from small variations in loudspeaker and microphone array positioning during IR measurement. As no simple method could be found to eliminate angular variations between stimuli, subjects were simply asked to ignore them.[26] Accordingly, subjects were instructed to make their similarity judgements on the basis of other attributes, and were explicitly asked not to refer to angular position in their comments.

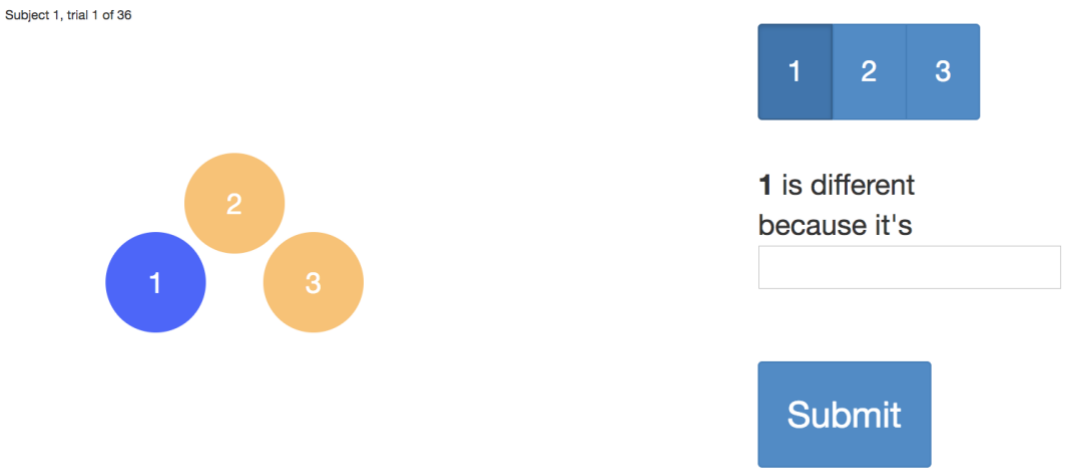


Figure 13 Triadic comparison interface

[26] Methods employed to minimize variations in perceived sound source angle included time alignment of the IR channels, to reduce inter-aural time delay cues, and loudness matching of the IR channels, to reduce inter-aural level difference cues. Though effective with many of the IRs, these techniques were not universally successful, for reasons that remain unclear.

3.2.1.4 Experiment structure

The experiment was designed to test whether low and high probabilities of each label model could be reliably discriminated. There were nine label models to evaluate, and each was tested with each of the four sound sources: chorus, drums, jazz voice and orchestra. The sound source and label model variables were fully crossed, resulting in a total of 36 experimental conditions (nine labels by four sound sources). Each subject was exposed to each condition exactly once, resulting in 36 trials per subject.

Within each trial, the distribution of low and high label probabilities was randomized. That is, subjects were presented either with two low probability stimuli and one high probability stimulus, or with one low probability stimulus and two high probability stimuli, with the choice made randomly.

As mentioned above, the stimuli used in the experiment were randomly sampled from the Spacebuilder library using a custom algorithm. Sampling was done without replacement, meaning that no stimuli were shared between subjects, and each stimulus that was evaluated, was evaluated exactly once. This meant that individual differences in perception or discrimination ability could not be investigated, but, on the other hand, meant that a greater number of stimuli could be studied. This was expected to yield results that would generalize better to the entire library of IRs. The sampling algorithm is detailed in Appendix E.

3.2.2 Results

As stated earlier, the test aimed to answer three questions about the label models. This section will first report which models succeeded in predicting perceptual differences. It will then identify and discuss trends in subjects' verbal descriptions of the label models. Finally, it will conclude by proposing a set of three perceptual attributes that seemed to explain most of the subjects' responses. In the subsequent section (3.3), these three attributes will be compared with the attributes in ISO 3382-1.

3.2.2.1 Which label models succeeded in predicting perceptual differences?

The first of the three research questions asked which label models were successful in predicting perceptual differences between stimuli. In each trial, three stimuli were presented, and one was predicted by a label model to sound different from the others. The subject was instructed to select the stimulus that they found the most distinctive. If the subject's choice agreed with the model's prediction, the trial was deemed a "success"; if the subject and the model disagreed, the trial was deemed a "failure".

A binomial test was used to determine whether each models' success rate was significantly different from chance. The results of this test are shown in Figure 14. A total of 60 trials were run for each label model. The number of successes is plotted on the y-axis. The number of successes expected by chance is shown by the solid line (1/3 or 20 trials); the number required to reject the null hypothesis that the models did not predict differences is shown by the dotted line ($\alpha = 0.05$). All models except *rich* gave better-than-chance results. It was concluded that the *rich* model was perceptually ineffective. Finding reasons for its failure was left for future work. The *rich* model was excluded from further analyses.

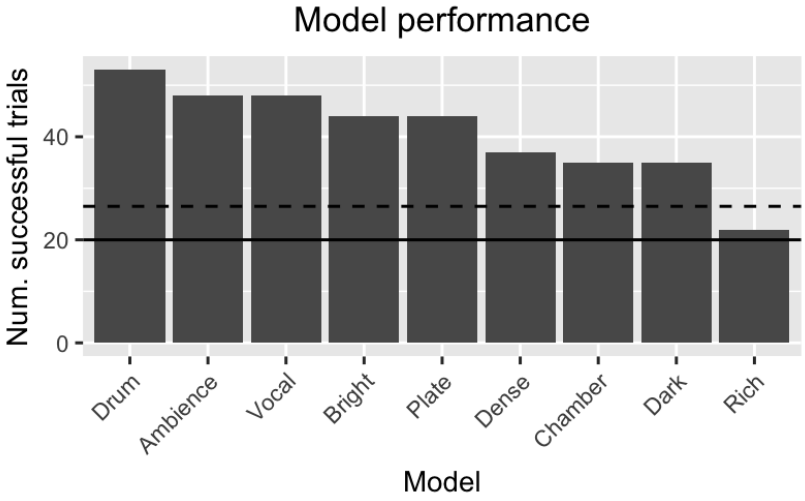


Figure 14 Model discriminability results

3.2.2.2 Trends in verbal descriptions

In addition to identifying the most perceptually different item within each stimulus trio, subjects were also asked to describe the nature of this difference with a word or short phrase. These verbal descriptions were collected for two reasons. First, an analysis of the descriptions elicited by subjects could shed light onto the most intuitive words for the putative attributes under investigation. It could be the case, for instance, that vocalness, as measured by *vocal* label probability, is indeed a genuine perceptual attribute of reverberation, but that differences in this attribute are most naturally described using a word other than "vocal". Perhaps differences in vocalness "sound like" differences "brightness" or "distance". Second, an analysis of subjects' verbal descriptions could provide evidence of correlations or dependencies among the nine attributes. It seems likely, for example, that the models for "bright" and "dark" describe the same attribute but with opposite polarities. If it were found that subjects used similar words to describe these two models, that would provide additional evidence that the two attributes might be (anti-)correlated, or related to the same underlying perceptual dimension.

This section will examine the descriptive words elicited by each label model. It will first explain some post-processing that was applied to subjects' responses to help elucidate trends. It will then examine the elicited terms. Finally, it will attempt to draw conclusions about the perceptual dimensions of natural reverb by examining relationships between verbal descriptions and signal features.

Note that the figures in this section show only the words from "successful" trials, where the subject and the model agreed on the most different stimulus.

3.2.2.2.1 Text response post-processing: converting raw responses to "attribute terms"

To help identify trends in the subjects' text descriptions of the "different" stimulus, the subjects' raw text responses were lightly processed. The goal of the processing was to extract the key idea from each response and to standardize the responses' spelling and wording. To distinguish between the unprocessed, original text responses, and the processed, standardized versions, two different designations are used. The subjects' original responses are referred to as *raw text responses*; this

processing step converted them into *attribute terms*. To transform the raw text responses into attribute terms, six rules were applied. The rules appear below. Examples of the six rules being applied are shown in Table 6.

Table 6 Examples of text processing rules

rule	raw text response	attribute term(s)
1	boxy, narrower bandwidth	boxy, narrow-bandwidth
2	honkier reverb (mid boost)	more-mids
3	bit longer, brighter maybe	long
4	longer reverb time	long
5	darker, stronger predelay	dark, more-predelay
5	larger early decay	more-early-decay
6	the one that sounds the most distant	far
6	Darker, not as spectrally extended	dark, not-full-range

The six rules were as follows.

1. If a text response seemed to refer to two separate perceptual attributes, it was converted into two separate attribute terms (e.g. the response “boxy, narrower bandwidth” became the terms “boxy” and “narrow-bandwidth”).
2. If a response attempted to explain a single attribute in multiple ways, the attribute was recorded only once. ("honkier reverb (mid boost)" became "more-mids")
3. If a response contained qualifying words that indicated uncertainly (e.g. “maybe”), the attribute associated with the qualifier was removed (“bit longer, brighter maybe” the “brighter maybe” part was removed).
4. Wherever possible, descriptions of the reverberant sound were reduced to simple adjectives (e.g. "long reverb" became simply "long"). This shorthand made the attribute terms more concise, and made sense since most ambiguous responses were assumed to refer to the reverberant sound.

5. Adjectives were normalized such that those indicating a greater magnitude of some attribute (e.g. “long(er)”, “strong(er)”, “larg(er)”) were mapped onto the single adjective “more”. Those adjectives indicating a smaller magnitude of an attribute were mapped to the adjective “less”.
6. Several specific less-frequently occurring terms were replaced by more common terms with similar meanings. Specifically, the term “far” was substituted for “distant”, “big” was substituted for “large”, and “not-full-range” was substituted for the phrase “not as spectrally extended”.

This post-processing was performed manually by the author. To avoid biasing the results, the task was done blindly. That is, while processing the responses, the author was not aware of which label was being evaluated, or of which subject had written the response.

3.2.2.2.2 Attribute terms elicited by each label model

Next, the attribute terms elicited by each model will be examined. For the purposes of this discussion, models will be grouped by the objective similarity of their outputs. Recall from Figure 11 that the predictions of the *drum* and *ambience* models were relatively highly correlated, as were the predictions of the *plate*, *bright* and *dark* models. The predictions of the *vocal*, *dense* and *chamber* models, on the other hand, did not strongly correlate with each other or with other models. Unsurprisingly, models whose predictions were correlated tended to elicit similar verbal descriptions. Each model will be discussed in turn, beginning with *drum* and *ambience*, continuing with *plate*, *bright*, and *dark*, and then concluding with *vocal*, *dense* and *chamber*.

Figure 15 shows the distribution of attribute terms collected from “successful” trials with the *drum* and *ambience* models. The plots are divided into two halves, with the left half showing the terms associated with low probabilities of the label, and the right half showing the terms associated with high probabilities of the label. Again, as discussed in section 3.2.1.4, each trial of the experiment consisted of either one low-probability stimulus and two high-probability stimuli, or two low-probability stimuli and one high-probability stimulus. The left side of each plot (low) shows terms elicited

in the former, low-high-high, types of trials; while the right side of each plot (high) shows terms from the latter, low-low-high, types of trials.

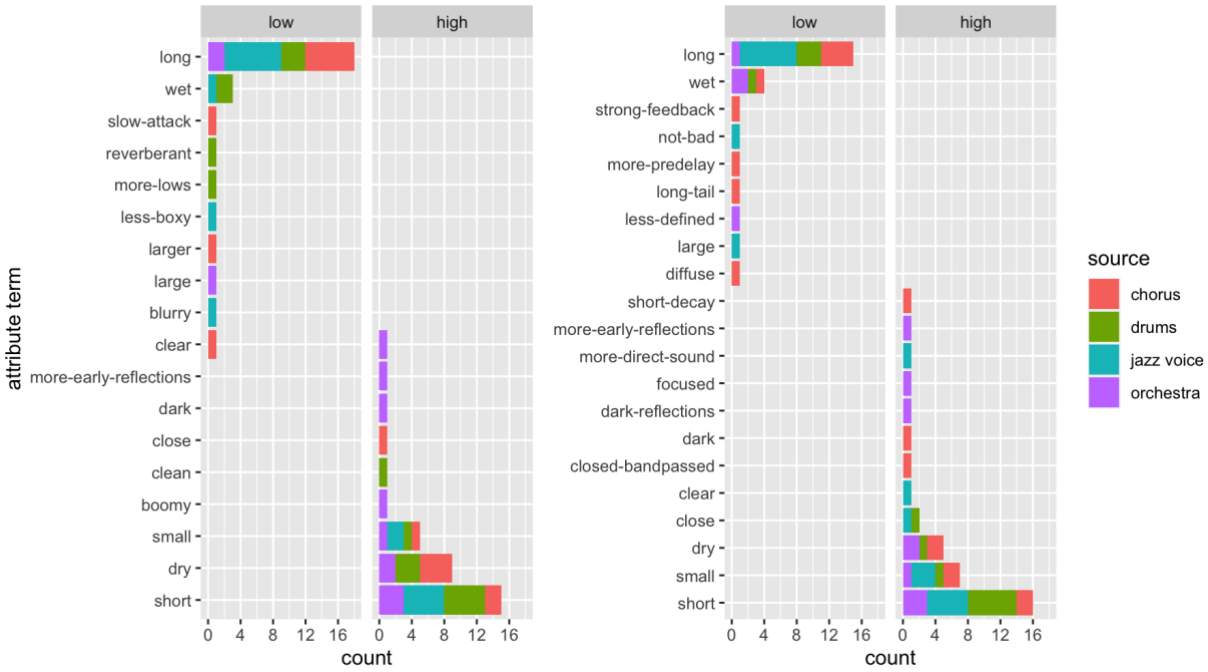


Figure 15 *Drum* and *Ambience* term distributions

Somewhat surprisingly, given the lack of semantic similarity between the labels "drum" and "ambience", the attribute terms elicited by the two models are quite similar. “Long” and “wet” are associated with low probabilities, while “dry”, “small” and “short” are associated with high probabilities. For both models, the single most common terms for the low and high probability conditions, respectively, are “long” and “short”.

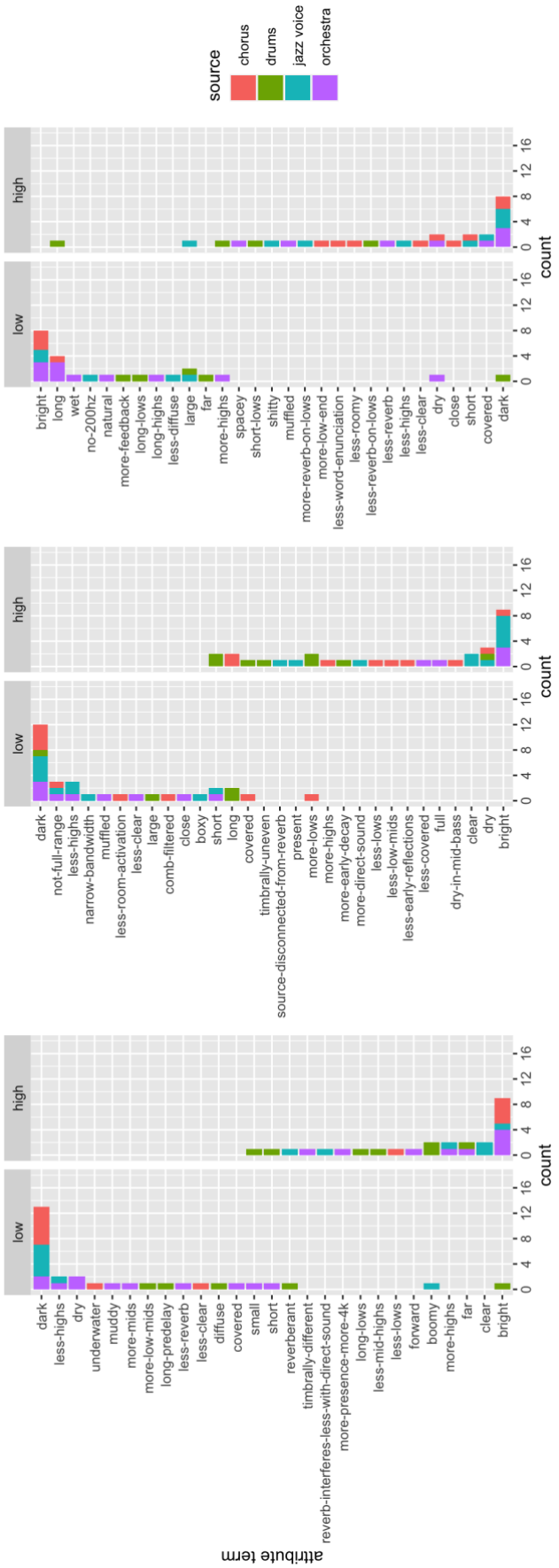


Figure 16 *Plate*, *Dark* and *Bright* term distributions

Figure 16 shows the distributions for the plate, bright and dark models, which also share many common terms. For bright and plate, the terms “dark” and “less-highs” are used to describe low probabilities, while “bright” is used to describe high probabilities. The terms for the dark model are similar but inverted. Intuitively, “dark” is used to describe high probabilities of *dark*, while “bright” is used to describe low probabilities.

The terms elicited by the *vocal* model are shown in Figure 17. “Short” and “long” are popular descriptions, but “close” and “far” also occur frequently.

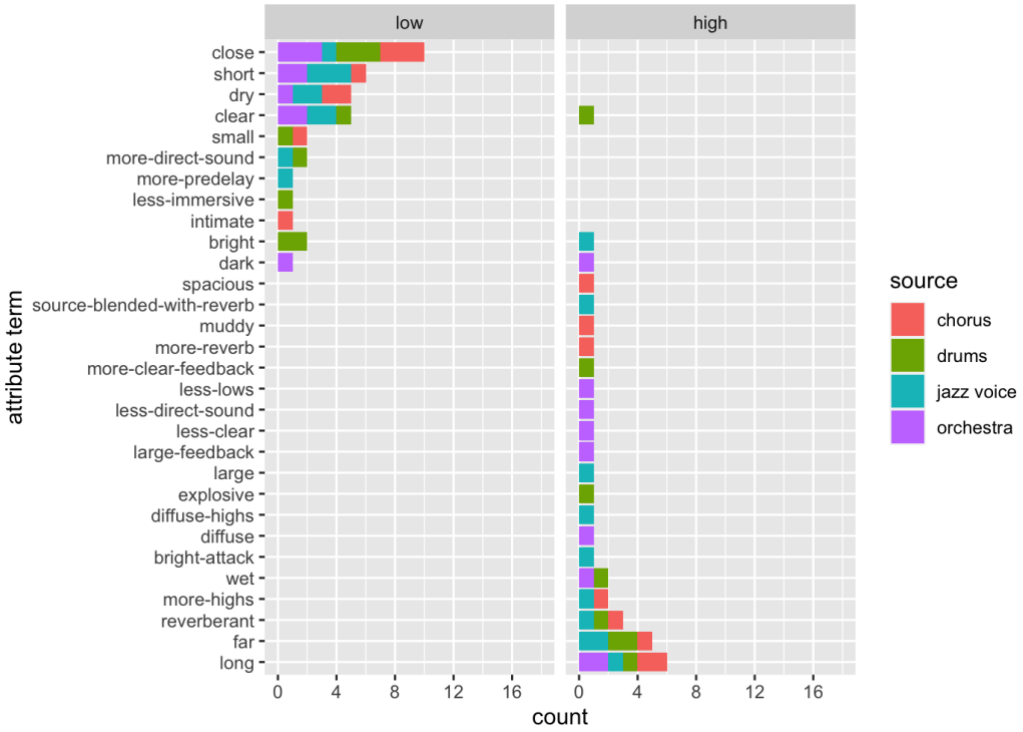


Figure 17 *Vocal* term distribution

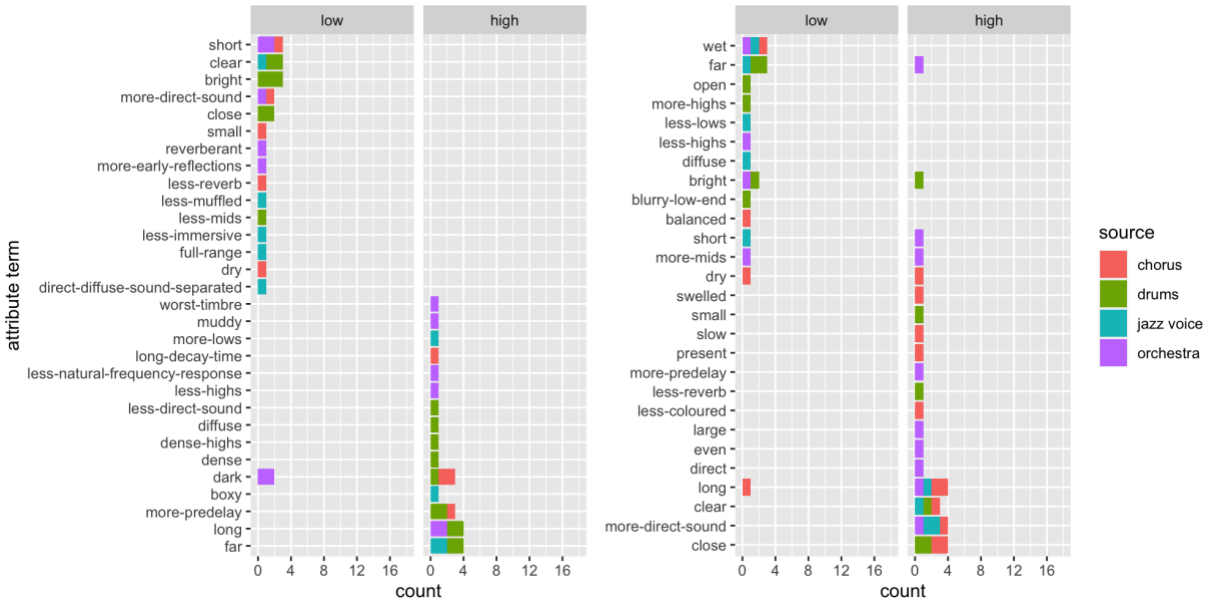


Figure 18 *Chamber* and *dense* term distributions

The distributions elicited by the *chamber* and *dense* models are shown in Figure 18. One notable feature of both is a larger dispersion of terms. While the models discussed earlier all had relatively spiky distributions, centered around a small number of popular terms, the dense and chamber distributions are much smoother and more spread out. Neither is clearly dominated by a single pair of words. As a result, these distributions are more difficult to summarize.

The next section will examine the relationships between these attribute term summaries and the signal features of the stimuli. Its goal will be to formulate a low-dimensional perceptual space that accounts for most of the perceptual variation observed in the listening test trials.

3.2.2.3 Perceptual attributes of natural reverb suggested by the listening test

Considering the term distribution summaries presented in the last section, subjects' verbal descriptions appear to be dominated by three pairs of terms: short/long, dark/bright and close/far. The pair short/long was associated with the *drum* and *ambience* models; dark/bright was associated with the *plate*, *bright* and *dark* models; close/far was associated with the *vocal* model. These similarities between the verbal descriptions of different groups of models, along with the knowledge that many of the models' predictions were correlated (c.f. section 3.1.3.1), suggest that the bulk of the perceptual variation heard between stimuli might be attributable to small number of underlying perceptual attributes. In other words, while this chapter had initially assumed that each model might be associated with a distinct attribute, the results suggest that these nine attributes - vocalness, darkness, plateness, brightness, density, richness, ambience-ness and drumness - may not, in fact, be different. Rather, the supposed attributes of ambienceness and drumness, may be more or less equivalent. Likewise, the attributes of darkness, plateness, and brightness, may be similar or equivalent, at least within collections of natural IRs.

Given that listeners are capable of making judgements about perceptual properties of sound sources and acoustic environments independently, for example distinct judgments about sound source width and room width (Rumsey, 2002; Rumsey & Berg, 2001), some consideration should be given to whether the elicited term pairs dark/bright and close/far refer to aspects of the auditory foreground (i.e., the sound source) or the auditory background (i.e., the reverberation). Some insight can be gained by examining subjects' raw text responses. In these short, freely elicited texts, subjects rarely specified to which auditory stream they were attending while making the judgement. A subject's response that the different stimulus was "brighter" could, in principle, have referred either to a brighter sound source or brighter reverberation. A minority of instances, however, did explicitly reference one stream or the other. With respect to the words "bright" and "dark", out of a total of 115 responses containing either word, 11 referred explicitly to the background stream (e.g. "brighter reverb") while only one referred to the foreground stream ("darker source"). Conversely, with respect to "close" and "far", out of a total of 39 responses, 18 referred explicitly to the foreground (e.g. "closer source") and none referred to the background. This suggests that while the trials in the experiment may have contained variations in both sound source brightness and

reverberation brightness, differences in reverberation brightness were likely heard more often. Likewise, while trials may have contained variations in sound source distance and reverberation distance, sound source distance variations appeared to be more common. It will be assumed, then, that when subjects referred to differences in brightness they generally meant differences in *reverberation brightness*, and that when they referred to differences in distance they generally meant differences in *sound source distance.*

The three underlying attributes that appear to account for most of the perceptual variation in the listening test are summarised in Table 7. Here, the "low term" and "high term" columns give the specific terms that were associated with low and high amounts of the attribute. The "implied attribute" column proposes names for the three attributes: *reverberance, (reverberation) brightness* and *source distance.*

Table 7 Term pairs and associated implied attributes

Associated label models	Low term	High term	Implied attribute
drum, ambience	short	long	*reverberance*
plate, bright, dark	dark	bright	*brightness*
vocal	close	far	*source distance*

3.2.2.3.1 Associations between implied attributes and signal features

Next, the relationships between these three implied attributes and the signal features of the stimuli were explored through a statistical analysis. The analysis sought to identify which features were most strongly associated with each attribute. It considered all trials that elicited one of the six attribute terms in question - short, long, dark, bright, close and far - regardless of which label model the trial was intended to test.

Prior to the analysis, the triadic comparison in each trial was first decomposed into a set of two pairwise comparisons. That is, within a trial, if a subject described stimulus

1 as “bright”, this was assumed to mean that it had more brightness than either stimulus 2 or stimulus 3. In the implicit pairwise comparison between stimuli 1 and 2, stimulus 1 is considered to have more brightness and stimulus 2 to have less brightness; likewise for the pairwise comparison between stimuli 1 and 3. This step converted the triadic comparisons in the listening test into a larger set of pairwise comparisons, where each item in each pair was labelled as having “less” or “more” of one of the three attributes.

Next, a paired *t*-test was performed on each combination of the three attributes and the 36 features listed in Table 4. The *t*-tests produced two outputs: a *t*-statistic, which described how strongly the signal feature was associated with relative differences in the attribute, and a *p*-value, which gave the probability of the associated *t*-statistic under a null hypothesis. The results of the *t*-tests are shown in Table 8. Only those tests with *p*-values less than 0.001 are shown. To highlight the strongest associations, the four largest *t*-statistics for each attribute are printed in boldface. The procedure used to create the table is explained in more detail in Appendix D.3.

To summarize the table, the implied attribute of reverberance appears to be correlated with the *EDT* and T_{30} features and anticorrelated with C_{80}. The implied attribute of brightness appears to be associated with Soulodre and Bradley’s *early bass level (EBL)* and *late treble ratio* (TR_{late}), as well as two wet features: *spectral slope* and *spectral skew*. Source distance appears anticorrelated with most bands of the *C80* and correlated with the late mid-frequency *number of peaks (NP Late 1-4k).*

Table 8 Paired t-test results for attribute/feature combinations

	Reverberance	Brightness	Distance
EDT 63-1k	**18.69**		
EDT 2-4k	**22.44**	7.24	5.58
EDT 8k	7.07	4.50	10.25
T30 63-500	14.54		
T30 1-2k	**14.67**	4.56	
T30 4-8k	12.13		5.52
C80 125	-7.38		-5.13
C80 250-2k	**-15.89**		**-12.43**
C80 4k	-9.00		**-13.63**
C80 8k	-4.95		**-12.59**
IACC early			-4.70
IACC late	-12.44	-3.58	
BR (EDT)		-3.74	
BR (T30)	-4.06	-4.55	
TR (EDT)	-8.10	4.89	4.76
TR (T30)	-5.92	5.45	6.11
EBL (SB)	-8.12	**-13.45**	7.81
TR (SB)		**13.87**	
NP early 63-125		-12.66	9.03
NP early 250-500	-5.14	-8.23	9.91
NP early 2k-4k		-5.26	11.17
NP late 63-125		-11.96	9.41
NP late 250-500		-7.63	10.69
NP late 1-4k		-4.69	**12.38**
NP late 8k			7.42
NED Mixing Time		7.22	
Log Attack Time		-3.64	6.22
Curvature			-3.56
pRev	13.26		
pClar	-10.98		
pASW	5.36		
pLEV	12.47		
Spectral Slope	-6.22	**14.57**	
Spectral Skew	5.58	**-14.88**	
Spectral Decrease		-6.23	-4.59
Spectral Flatness	-9.63	4.65	

3.3 Discussion: contrasting experimental attributes with ISO 3382 attributes

This section will compare the three attributes suggested by the experiment with the four attributes in the ISO model. Its aim will be to identify any dimensions of perceptual variation that aren't accounted for in ISO 3382-1. The specific words used by subjects to describe these attributes will also be considered, in particular with respect to whether they reveal any particularly intuitive labels for search interface controls.

The three implied attributes will be treated in the order they were introduced: first reverberance, then brightness, then distance.

3.3.1 Reverberance

An examination of Table 8 shows that implied attribute of reverberance is strongly associated with the *early decay time (EDT)* feature. *EDT* is also known to correlate with the ISO attribute of perceived reverberance. This suggests that the experimental attribute and the ISO attribute are synonymous. The data give no reason to suspect that the reverberance attribute revealed in the experiment is any different from the similarly named attribute in the ISO 3382 model.

That said, the words chosen by the experimental subjects, all trained sound engineers, to describe variations in this attribute may give insight into how best to present controls for it in an audio effect interface. Specifically, even though the academic literature refers to this attribute as "reverberance", the specific word "reverberance" did not often occur in the elicited raw responses: only in 16 cases was a stimulus described as more or less "reverberant" than the others in a trial. Words loosely related to reverberance that appeared much more often included the pairs large/small (n=36), wet/dry (n=54)

and long/short (n=176).[27] The frequencies of these elicited words suggest that terms describing the duration of the subjective decay (e.g. long/short) may be more intuitive for sound engineers than terms related to reverberance levels. A widget controlling reverberance in a reverb plugin might be slightly more easily understood if it were labelled "decay duration" or "decay time" rather than "reverberance level".

3.3.2 Source distance

The implied attribute of sound source distance raises interesting questions about both subjects' use of language and acoustic reverberation's perceptual structure.

Concerning language, at first glance the strong relationship between distance and the C_{80} feature suggests that this attribute may be perceptually similar, or even identical, to the ISO 3382 attribute of perceived clarity, also associated with C_{80}. If the two attributes of source distance and perceived clarity are the same, this would mean that during the experiment, subjects heard differences in perceived clarity but instead of using clarity-related words to describe these differences (e.g. "more clear", "less clear"), chose instead to use words related to distance (e.g. "closer", "farther"). The raw text responses certainly suggest an inclination toward distance-related words. Within this data, references to distance (i.e. close/distant/near/far) outnumber references to clarity (clear/clarity) by a ratio of almost 2:1 (n=55 vs n=35).

If it were true that subjects spontaneously chose distance-related words to describe differences in perceived clarity, this would have implications for IR search interface design. It would suggest that distance-related words might be more intuitive descriptions for variations in C_{80} than clarity-related words. A more intuitive description might make the meaning of a C_{80}-based control easier to understand during a user's first encounter with an interface. That is, using words such as "closer" and

[27] These numbers were calculated by examining subjects' raw text responses and counting the number of answers that included either word in the pair. Thus, 176 responses included either "long" or "short", 54 included "wet" or "dry", and 36 included "large" or "small". Sixteen responses included the word "reverberant" (sometimes misspelled), along with modifiers such as "more", "too", "most" and "less".

"farther", rather than "more clear" and "less clear", on a C_{80} control might make an interface easier to learn.

On the other hand, it is also possible that the implied attribute of source distance and the ISO attribute of perceived clarity are not perceptually identical. If this were true, it would have implications for perceptual models of reverberation. Namely, it would suggest that a fifth attribute, distinct from perceived reverberance, perceived clarity, ASW and LEV, could be useful for understand perceived differences between reverb outputs. Independence between the attributes of source distance and perceived clarity would imply that it was possible for sound source images in natural reverberation to appear simultaneously both "close" and "unclear" as well as both "far" and "clear". If the attributes were distinct, it would also mean, in theory, that they could be best predicted by different signal features. Since perceived clarity is well predicted by C_{80}, this would mean that some other signal feature would be a better predictor for distance. In this case, the strong relationship observed in the experiment between source distance and C_{80} would not be due to C_{80} being the best possible predictor for distance, but rather only the best from the set of features considered.

3.3.3 Brightness

Of the three perceptual attributes suggested by the experiment, the one least likely to have a correlate in ISO 3382 would seem to be brightness. Nonetheless, brightness is clearly an attribute of interest to mixing engineers, as brightness-related words are common in algorithmic preset names (c.f. Figure 8). Brightness, then, appears to be a genuine attribute of reverberation that is not present in the ISO model.

The words used by subjects to describe brightness differences are unsurprising. More interesting, however, are the signal features associated with brightness variations, as revealed in Table 8. To predict timbral attributes such as brightness, the room acoustics community has traditionally used IR-based features such as the *early bass level (EBL)* and *late treble ratio (TR_{late})* (see section 2.2.2.2). As expected, these features had strong relationships with brightness in the experiment. However, even stronger relationships were found between brightness and two wet features: *spectral slope* and

spectral skew. These features originate in the musical timbre perception community (Peeters et al., 2011) and have never, to the knowledge of the author, been applied in the context of reverberation assessment. These strong associations suggest that wet features summarizing spectral shape may prove to be better predictors of reverberation brightness than IR features, an intriguing but unexplored proposition.

3.4 Summary

This chapter set out to investigate the perceptual structure of natural reverberation. This was accomplished by first identifying a set of putative reverberation attributes suggested by the names of algorithmic reverb presets. For nine descriptive labels appearing in preset names, statistical models were built that related the label to a set of signal features. Each label model corresponded to a putative attribute of reverberation.

A listening test was then conducted to determine whether any of these putative attributes could explain perceptual differences within a library of natural IRs. The speculative attributes were compared with the four in the ISO 3382 model, to see if any represented distinct sources of perceptual variation not accounted for in ISO 3382-1. To this end, two conclusions can be drawn. One is clear, and one is speculative.

The clear conclusion is the existence of the attribute of reverberation brightness. Brightness explains many of the differences reported in the experiment, and it has no correlate in ISO 3382. With respect to the question of which signal features are most predictive of brightness, the results suggest that the wet features *spectral slope* and *spectral skew* might be more effective than the more traditional IR-based features *early bass level* and *late treble ratio*. This proposition will be considered further in the next chapter.

The more speculative conclusion concerns the attribute of sound source distance. The prospect of a distance attribute arose from the relative frequency of the terms "close" and "far" in subjects' elicited responses. Physically, "close" and "far" were associated with variations in the C_{80} feature, itself known to correlate with the ISO attribute of perceived clarity. This suggests either that 1) the attribute of source distance is identical

to ISO perceived clarity, and that the words “close” and “far” are merely intuitive verbal descriptions of low and high levels of this attribute, or that 2) source distance is perceptually distinct from ISO clarity, and, though correlated with C_{80}, could be better predicted by some other feature not tested in this chapter. One type of evidence that would lend credence to proposition 2) would be the identification of a signal feature that was a better predictor of source distance judgements than C_{80}. If such a feature were found, this would suggest that the two attributes were distinct, and that source distance, like reverberation brightness, was another attribute of acoustic reverberation not accounted for the ISO model. The question of the independence of source distance and perceived clarity will also be explored in the next chapter.

4 MODELS OF SOURCE DISTANCE AND REVERBERATION BRIGHTNESS

4.1 Introduction

The investigation in the previous chapter identified two attributes as important sources of perceptual variation within a library of natural impulse responses. One attribute was associated with the words "bright" and "dark", while the other was associated with the words "close" and "far". These two attributes were given the names *brightness* and *source distance*.

If these two attributes are indeed salient perceptual dimensions of natural IR libraries, it follows that a perceptual search interface should enable searches along them. Enabling such searches, however, requires that levels of brightness and source distance be assigned to each IR. Assigning levels at scale requires automatic, computational techniques.

This chapter aims to develop such computational techniques. Specifically, it compares the predictive ability of a set of simple computational models of each attribute. Each model is a function that accepts a single signal feature as input and, as output, produces an estimate of the subjective rating of one of the attributes. The findings of the investigation will be used to propose predictive models of the two attributes to be used in a search interface.

This chapter will first detail the candidate models to be compared. This will be followed by a methodology section, in which the method used to collect perceptual judgements will be explained. Following this, a results section will examine each model's performance. Differences between the models will be considered, and, finally, the two models that appear best able to predict the two attributes will be presented. The chapter will conclude with a discussion of the relationship between the attributes of source distance and ISO perceived clarity, to address questions raised in the previous chapter.

4.2 Candidate models and hypotheses

As will be detailed further in section 4.3.3.2, all of the candidate models in this chapter have a simple mathematical structure: they consist of one signal feature transformed by a sigmoid function. Because one feature is so integral to each model, in the sections below will sometimes use the terms "feature" and "model" interchangeably. In a strict sense, however, a "feature" is a number calculated from a signal, and a "model" consists of a signal feature transformed by a sigmoid function.

This section will first outline the candidate models for the source distance attribute, and then the candidate models for the brightness attribute. At the end of each section, specific hypotheses about the performance of each candidate model will be listed.

4.2.1 Source Distance

The features used in the candidate models for source distance fall into two groups. The first group contains features designed to predict ISO perceived clarity, or the balance between early- and late-arriving sound. The features in this group are drawn from the literature on room acoustics perception. The second group contains two features explicitly designed to predict source distance. These features are drawn from the more general literature on computational perception. Each group will be discussed in turn.

4.2.1.1 Clarity-related features

In the previous chapter, several features were shown to be particularly strongly associated with our implied source distance attribute. These were the C_{80} *250-2k*, C_{80} *4000* and C_{80} *8000*. A correlation between variants of the C_{80} and perceived source distance is unsurprising. Even though C_{80} is intended to measure signal intelligibility, in practice it functions by calculating a ratio of early- to late-arriving sound. This ratio is conceptually similar to the direct-to-reverberant ratio (DRR), which is known to be an important cue to auditory source distance (Zahorik et al., 2005).

The "clarity-related features" investigated in this chapter will be the three drawn from the previous chapter (C_{80} *250-2k*, C_{80} *4000* and C_{80} *8000)*, as well as a C_{80} *average* (an average of the 125-8000 Hz bands), and the *pClar* feature. *pClar* also attempts to calculate a ratio of early- to late-arriving sound, but, unlike the C_{80} , does so from a wet signal, using a binaural model (see section 2.2.2.4).

4.2.1.2 Auditory source distance-related features

As mentioned previously, a primary cue to source distance is thought to be the direct-to-reverberant ratio (DRR): the ratio between the sound energy assigned by the auditory system to the sound source, and the sound energy assigned by the auditory system to the acoustic environment.[28] As the methods used by the auditory system to discriminate direct from reverberant sound are not fully understood, computing an exact perceived DRR is not currently possible. Several heuristic methods to estimate the DRR have been proposed, however, two of which are employed in this chapter. The first method, used by a novel feature called the $C_{10}{}'$, relies on temporal characteristics of the IR. The

[28] Distance judgments in the absence of this cue, for example judgements made in anechoic environments, tend to be exceptionally inaccurate, as reported for example in Nishimura and Sasaki (2004).

second method, used by a feature called the *ECDRR*, or *Equalization-Cancellation Direct-to-Reverberant Ratio*, relies on binaural differences in the wet signal. [29]

4.2.1.2.1 $C_{10}{}'$

The $C_{10}{}'$ feature estimates the DRR by applying a tapered window of about 10 ms to the beginning of an IR. Energy falling within the window is considered "direct", while energy outside the window is considered "reverberant". The ratio of these values, raised an exponent, produces a subjective distance estimate. The feature is shown in Equation 15; the window used to produce direct and reverberant estimates $\hat{D}$ and $\hat{R}$ is shown in Figure 19.

$$C_{10}{}' = \left(\frac{\hat{R}}{\hat{D}}\right)^{j}$$

Equation 15 The $C_{10}{}'$ feature

[29] A third heuristic for distinguishing direct from reverberant sound, which was investigated but ultimately not employed in this work, draws on theories of Griesinger and concerns phase coherence in reverberant tonal signals (2015). Griesinger suggests that phase alignment of harmonics in the region of 1.5 to 4 kHz serves as a cue used for foreground/background stream segregation and hence, potentially, for direct-to-reverberant ratio and distance estimation. Although some evidence supports his hypothesis (Lokki et al., 2011), no preexisting signal analysis code to quantify phase coherence could be located. As such, determining the utility of this phase coherence cue for estimating perceived source distance was, regrettably, left for future work.

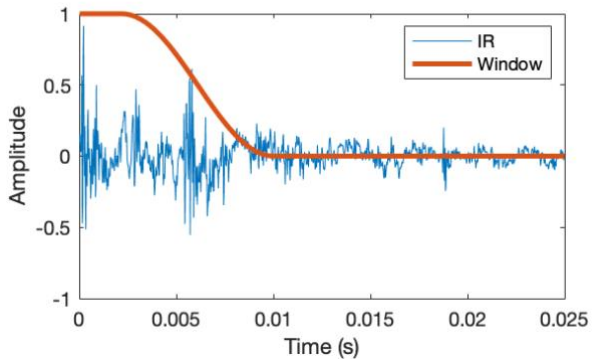


Figure 19 Temporal window for $C_{10}{}'$

The $C_{10}{}'$ feature is inspired by the work of Bronkhorst and Houtgast (1999), who predicted source distance as a function of an impulse response and a room's reverberation radius. Their original model is shown in Equation 16, where d is a perceived distance, r is the reverberation radius, A and j are constants, and $\hat{D}$ and $\hat{R}$ are estimates of the direct and reverberant energies, derived using the temporal window shown above.

$$d = Ar\left(\frac{\hat{R}}{\hat{D}}\right)^{j}$$

Equation 16 Bronkorst and Houtgast's distance model

The $C_{10}{}'$ feature used here differs from Bronkhorst and Houtgast's model in two ways. First, it omits the reverberation radius term, r, as this information was not readily available for the IRs in the library. Second, it omits the constant A, as this value was calculated implicitly by our fitting procedure (see section 4.3.3.2). For the constant j, the same value proposed by Bronkhorst is used (0.49). Likewise, $\hat{D}$ and $\hat{R}$ are calculated via the same tapered window.

4.2.1.2.2 ECDRR

Unlike the C_{10}', the *ECDRR*, or *Equalization-Cancellation Direct-to-Reverberant Ratio*, operates on the wet, binaural signal, rather than the IR (Lu & Cooke, 2010). It is inspired by research into binaural masking effects (Durlach, 1960). The *ECDRR* functions by first estimating the angle of arrival of the direct sound. This angle is determined through a cross-correlation process on the outputs of a binaural auditory filterbank. Once the angle of the direct sound has been estimated, the energy arriving from this direction is removed from the signal via an "equalization-cancellation" procedure. The energy that is successfully removed is considered to be "direct", while the remaining energy is considered to be "reverberant". The *ECDRR* is the ratio of these two quantities.

In the experiment, the *ECDRR* feature was calculated using MATLAB code provided by the feature's creators. As in Lu and Cooke's original work, analysis windows of 200 ms with no overlap were used. One *ECDRR* value was produced for each window of the wet signal; the average over all windows was taken to be the feature's final value.

4.2.1.3 Hypotheses

Of the seven features mentioned above, only two, the *ECDRR* and the C_{10}', are explicitly designed to predict source distance rather than perceived clarity. Subjects in the experiment to follow will be explicitly asked to rate distance. If it is true that source distance and perceived clarity are independent attributes, these distance-specific features might be expected to better correlate with the collected distance judgements. In particular, as the *ECDRR* operates on a wet signal, in fact the exact stimulus presented to subjects, rather than simply on the impulse response used to generate the stimulus, the *ECDRR* feature would be expected to perform best overall.

Concerning the five clarity-related features, it seems reasonable to suppose that all would be moderately successful at predicting source distance. Those features that operate only on high-frequency information, namely the C_{80} *4000* and C_{80} *8000*, might be expected to perform slightly less well, as they fail to take into account distance cues in the mid-range and lower frequencies. With respect to the remaining features, *pClar*,

C_{80} *average* and C_{80} *250-2k*, as little evidence suggests any one of them to be dramatically better than the others at predicting perceived clarity, no large differences are expected in their ability to predict the related attribute of source distance.

4.2.2 Brightness

The features used in candidate models for brightness fall into three groups. The first contains two traditional features derived from the IR: the *early bass level* and *late treble ratio*. The second group contains features that operate on the wet signal: *spectral slope (wet)*, *spectral skew* and *spectral decrease*. The third group contains only a single feature, the *spectral slope (IR)*. *Spectral slope (IR)*, like the features in group one, operates on the IR signal. Unlike the features in either groups one or two, however, the *spectral slope (IR)* was designed by the author, rather than being drawn from the existing literature. Each group will be discussed in turn.

4.2.2.1 Group one: early bass level *and* late treble ratio

As reviewed in section 2.2.2.2, the room acoustics literature includes a number of signal features purported to correlate with timbral attributes of reverberation. Two of the most successful have been Soulodre and Bradley's *early bass level (EBL)* and *late treble ratio* (TR_{late}) (1995). These two features were originally proposed as objective correlates of two separate perceptual attributes: the perceived levels of low frequencies in the signal and the perceived levels high frequencies in the signal, respectively. Despite having been designed for distinct purposes, however, both were associated quite strongly with differences in the brightness attribute discussed in the last chapter.

4.2.2.2 Group two: spectral slope (wet), spectral decrease and spectral skew

The last chapter showed that several wet features had even stronger relationships with the implied attribute of brightness than either the *EBL* or TR_{late}, however. These promising wet features were *spectral slope (wet)*, *spectral decrease* and *spectral skew*.

These features all describe trends in the shape of the wet signal's spectral envelope. Spectral envelope shape has had a long association with brightness attributes in musical instrument timbre research, in particular via a fourth feature called the *spectral centroid* (e.g., Schubert et al., 2004). Although the *spectral centroid* is not explicitly used in any of this chapter's candidate models, it was calculated on the stimuli and found to be almost perfectly correlated with the *spectral slope (wet)* (r=1)[30] and highly anti-correlated with the *spectral skew* (r =-.86). The high correlation between the *spectral centroid* and two of the wet features, as well as the conceptual similarity of all three wet features, suggests that all three are strong candidates for predicting brightness in the context of natural reverberation.

In this chapter, these three features were calculated with the Timbre Toolbox (Peeters et al., 2011). The "ERB fft" auditory model was used; all other analysis parameters were left at their defaults, including the window hop size (5.8 ms). The median value over all signal windows was taken as the final value of each feature.

4.2.2.3 Group three: spectral slope (IR)

While the three wet features, and the *spectral slope (wet)* in particular, would appear to be promising predictors for brightness, wet features are somewhat inconvenient to use in the context of IR search interface design. This is because wet features are calculated on the output of a convolution between an IR and a sound source, and thus require knowledge of the sound source being auditioned. Typically, in audio production contexts, the sound source (i.e., the stem or live signal to which the mixing engineer wishes to apply reverb) will only be chosen during the mixing session and is not available beforehand, when the signal features are most easily computed.

In an effort to find a quantity that was similar to the *spectral slope (wet)*, but that could be computed without knowledge of the sound source, a novel feature was proposed called the *spectral slope (IR)*. As the name implies, this feature is calculated from the IR signal alone.

[30] The exact correlation found was 0.99998, which is rounded up in the text to 1.

Specifically, the *spectral slope (IR)* is defined as the slope of the line of best fit through an IR's smoothed magnitude spectrum. [31] Smoothing is performed via a moving average along the frequency axis, using windows 1/3 octave wide and with a hop size of 1/12 octave. The spectrum is plotted on logarithmic axes for both frequency and amplitude. Frequencies from 63 Hz to 12500 Hz are considered, and only the first 300 ms of the IR is analyzed.

The implementation details of the *spectral slope (IR)* are similar to, and directly inspired by, the *deviation of level* feature of Takahashi et al. (2008).[32] The two features differ only with respect to how they summarize the smoothed spectrum. The *deviation of level* takes the standard deviation of smoothed spectrum, and thus measures the spread of its values, while the *spectral slope (IR)* measure its slope. A code listing for this feature is given in appendix B.1.

4.2.2.4 Hypotheses

While the *early bass level* and *late treble ratio* have shown high correlation with two perceptual attributes described as "the strength of the lows relative to mids" and "the strength of highs relative to mids", respectively (Soulodre & Bradley, 1995), and while these attributes may be similar, or even identical, to the attribute of brightness, neither feature has been shown to correlate strongly with brightness ratings specifically. Also, these features take into account only the IR, rather than the complete wet signal presented to the listener, and so are unable to consider any interaction effects with the sound source.

The three wet features, on the other hand, do take into account sound source effects. Further, the wet features, especially the *spectral slope* and *spectral skew*, have strong mathematical and conceptual similarities with another feature, the *spectral centroid*, which has been shown to predict brightness well in musical timbre contexts. For these reasons, the three wet features, especially the *spectral slope* and *spectral skew*, are expected to predict brightness judgments more effectively than the *early bass level* or the *late treble ratio*.

[31] The slope is equivalent to the *b1* term of an ordinary least-squares regression.
[32] See the earlier discussion of features related to spectral irregularity (section 2.2.2.2).

The novel feature, *spectral slope (IR)*, being a type of feature traditionally associated with brightness, but also being unable to capture sound source effects, is expected to have a middling performance level, in between that of the *early bass level* and *late treble* ratio, and that of the wet features.

4.3 Methodology

In this section the methods used to collect and model subjective judgements of brightness and source distance are explained. Judgments were elicited through a listening test. The subject pool and stimuli will be discussed first, and the test's user interface will then be described. The general mathematical structure of the candidate models will then be presented, followed by the metric used for model comparisons.

4.3.1 Experimental subjects

Seventeen subjects were recruited from the Sound Recording Area of McGill University's Schulich School of Music. All subjects were graduate students or faculty members at the School, and all reported at least two years of audio engineering experience. Distributions of subject age, academic level and audio engineering experience are shown in Figure 20. The labels "SR1" and "SR2" refer, respectively, to the first and second year of a 2-year master's degree program.

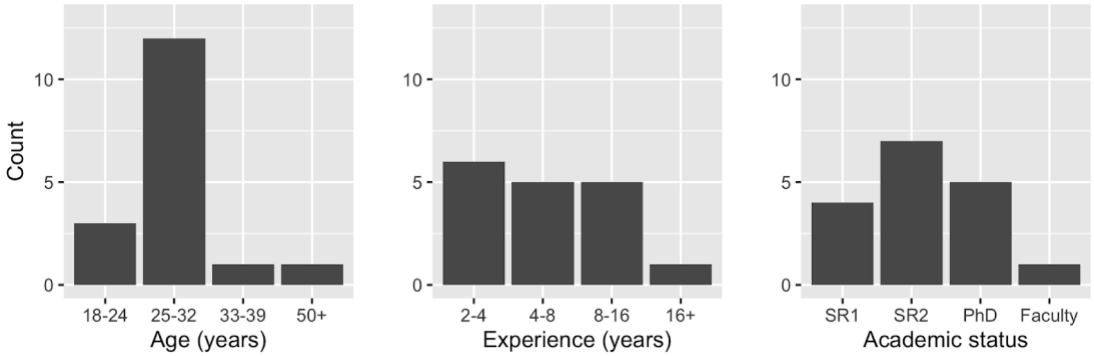


Figure 20 Experimental Subject Demographics

4.3.2 Experimental stimuli

All stimuli used in the test consisted of a "100% wet" reverb signal, i.e. the output of a convolution between a stereo IR and a monophonic sound source. Two different sound sources were used: a solo jazz voice recording and a drum kit recording (see Appendix C).

All stimuli were loudness equalized to -18 LUFS (International Telecommunications Union, 2011).

4.3.2.1 IRs

The IRs used in the experiment were drawn from a large subset of the Spacebuilder library (see Appendix A). To create stimuli, IRs were sampled randomly from the subset, subject to the constraint that each experimental block contained seven unique IRs.

4.3.3 Experiment design

The experiment was structured as a series of blocks. Within each block, subjects compared seven stimuli. Some blocks were designed to elicit judgements of brightness, while others were designed to elicit judgements of source distance.

In brightness blocks, subjects were asked to report the brightness of the stimuli, both in relation to each other, and in relation to two anchor points. The anchor points were indicated by short black lines on the rating scale. These anchor points were explained as the brightness levels corresponding to the subject's internal concept of "very dark" and "very bright" reverb. The verbal instructions thus implied that subjects should attend to the brightness of the auditory background stream (i.e., reverberation) rather than the brightness of the foreground stream (i.e., the sound source). Subjects recorded their ratings by positioning circular icons along a scale. Icons could be positioned using either a trackpad, or the left and right arrow keys on a keyboard. A screenshot of the brightness interface is show in Figure 21.

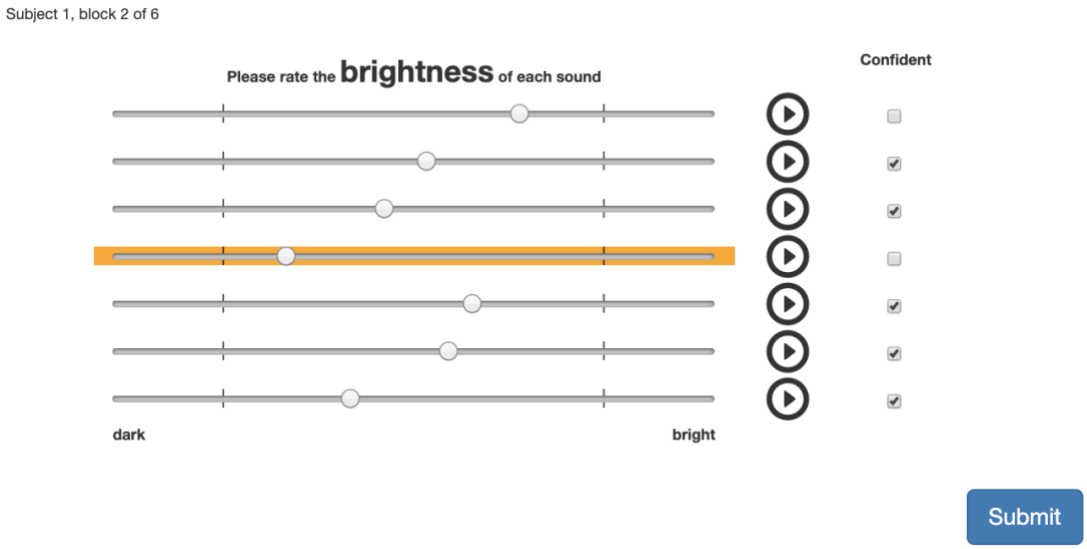


Figure 21 Brightness interface

Additionally, subjects were asked to indicate their confidence in their ratings using a check box. The box was "checked" by default, and subjects were asked to "uncheck" it if a stimulus' brightness was ambiguous or hard to characterize. This confidence data was not analyzed in the current investigation.

Distance blocks were largely analogous to brightness blocks. In these blocks, subjects were asked to indicate the perceived egocentric distance of the sound source, both in relation to each other, and in relation to two anchor points. The low anchor indicated a stimulus that sounded "inside [the subject's] head", while the high anchor point indicated a stimulus that had a perceived distance of "half a football field away". These verbal instructions thus implied that the distance judgements should be made while attending to the foreground stream (i.e., sound source distance) rather than the background stream (i.e., reverberation distance). The interface for distance blocks is shown in Figure 22.

As before, subjects were asked to indicate their confidence in their ratings using a binary checkbox. In addition, subjects were also asked to indicate whether an auditory image was "more than 30 degrees off-center", and also to indicate whether they were unable to make a meaningful distance judgement because they couldn't localize a sound source. Only those stimuli in which subjects said they were able to localize a sound source were analyzed.

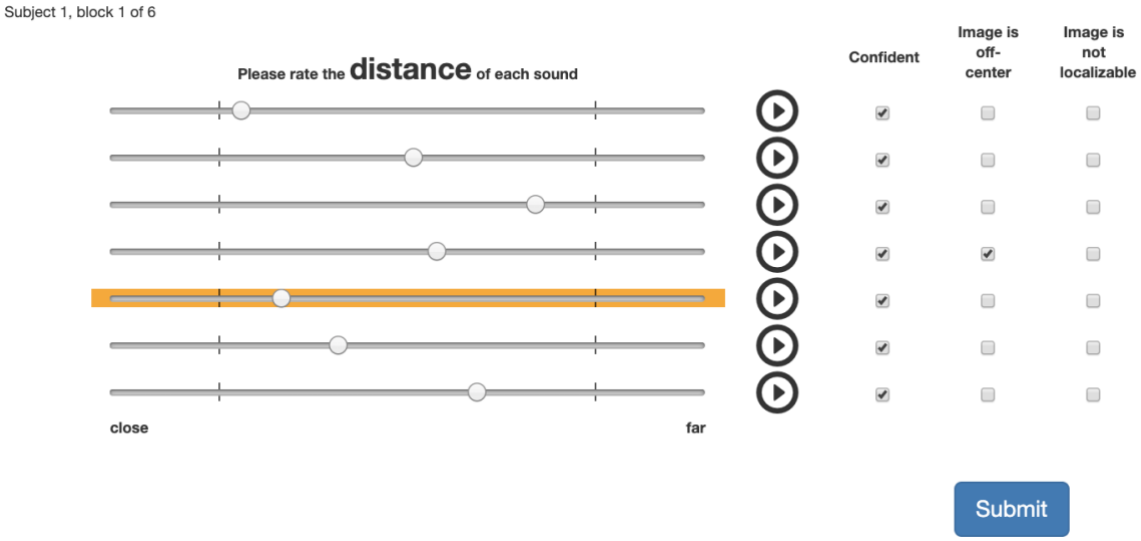


Figure 22 *Distance* interface

To enable rapid comparisons within each block, subjects could switch seamlessly between continuously playing stimuli using the trackpad or keyboard. At the end of each trial, before submitting their responses, subjects were asked to verify their answers by listening to the stimuli in sequence, in the order of their assigned rating (e.g. the stimulus rated least bright, followed by the stimulus rated next brightest, followed by the stimulus rated next brightest, etc.) Two keyboard shortcuts facilitated these sequential comparisons by cycling through the stimuli in order of increasing or decreasing attribute judgements.

Stimuli were auditioned over headphones. Subjects were able to control playback volume and were instructed to set it to a comfortable level.

4.3.3.1 Experiment length and block structure

The first few subjects to take the test completed a short version containing only six blocks. When it was found that six blocks could be completed relatively quickly, in about 20 minutes, it was decided to lengthen the experiment to eight blocks. Five subjects completed the short version and twelve completed the long version. The structures of the short and long versions are shown in Table 9 and Table 10. In both versions, the order of presentation of the blocks was randomized for each subject.

Table 9 "Short" block structure

Block number	***Attribute***	***Sound Source***
1	brightness	singing
2	brightness	drums
3	distance	singing
4	distance	drums
5	distance	singing
6	distance	Drums

Table 10 "Long" block structure

Block number	***Attribute***	***Sound Source***
1	brightness	singing
2	brightness	drums
3	distance	singing
4	distance	drums
5	distance	singing
6	distance	drums
7	brightness	singing
8	brightness	drums

To generate stimuli, IRs were randomly sampled from the library subset and then convolved with one of the two sound sources. Sampling was done without replacement for each sound source. This meant that each stimulus was rated only once by one subject; no stimuli were shared between subjects. This resulted in brightness ratings on a total of 406 stimuli and distance ratings on a total of 476 stimuli. The different numbers of stimuli for each attribute resulted from the fact that the "short" block structure, used by five subjects, contained fewer brightness blocks than distance blocks. In each case, half of the stimuli used the drums source and half used the jazz voice source.

4.3.3.2 Predictive model structure

In order to predict brightness and distance ratings from signal features, a transformation was required to map the domain of the features onto the range of the ratings. All collected attribute ratings lay on the unit interval, between 0 and 1, while the signal features used in the models took on a wide array of values. The C_{80} *average* feature, for example, extended on the test stimuli from about -8 to about 26, a range of 34 units, while the *spectral decrease* extended only from about 0.02 to 0.04.

The transformation from feature values to rating predictions was accomplished via a sigmoid function, as shown in Equation 17. All candidate models had this general structure, where x is a feature value, $\hat{y}$ is a predicted rating, and $\hat{\beta}_0$ and $\hat{\beta}_1$ are model parameters.

$$\hat{y} = \frac{1}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x}}$$

Equation 17 General model structure

4.3.3.3 Predictive model evaluation

To evaluate each candidate model, a sigmoid function was fit to the experimental data using the Gauss-Newton algorithm (i.e. using the Nonlinear Least Squares Regression function in R 4.1; R Core Team, 2021). That is, values of parameters $\hat{\beta}_0$ and $\hat{\beta}_1$ were found that minimized the *mean squared error* (MSE) between the model predictions $\hat{y}$ and the subjective ratings. This minimized MSE was also used as a metric to compare the models to one another. Models that predicted the ratings well had lower MSE, while models that predicted poorly had higher MSE.

4.4 Results

The experimental results, showing subjects' ratings, the models of best fit, and the best fitting models' MSEs, are presented here. Distance ratings appear in Figure 23 and Figure 24, and brightness ratings in Figure 25. Models are presented in order of increasing MSE, for the average of the two sound sources. The best fitting models appear at the tops of the figures.

4.4.1 Source distance

For the source distance attribute, the best fitting model for both sound sources was the C_{80} *average*. For jazz voice the MSE was 0.015 and for drums it was 0.020.

4.4.1.1 Outliers

Some outliers are evident, visually, in the *ECDRR* and $C_{10}{}'$ plots. Each model will be addressed in turn. Generally, these outliers can be attributed to atypical IRs, and idiosyncrasies in how the two models handle unusual inputs. At the end of this section, the experimental data will be reanalyzed with the outliers removed. A comparison of the results with and without outliers shows that these few values do not change the ranking of the models or the main conclusions of the study.

4.4.1.1.1 ECDRR

One sole IR can be seen to have an extremely large value of *ECDRR*. This value is 235, a full 15 standard deviations from the mean (1.8). An examination of the IR in question found that its two channels were nearly perfectly correlated (r=0.9952). In other words, the IR was nearly monophonic. Further investigation found that this high correlation between the left and right channels was caused by human error during the IR measurement process.

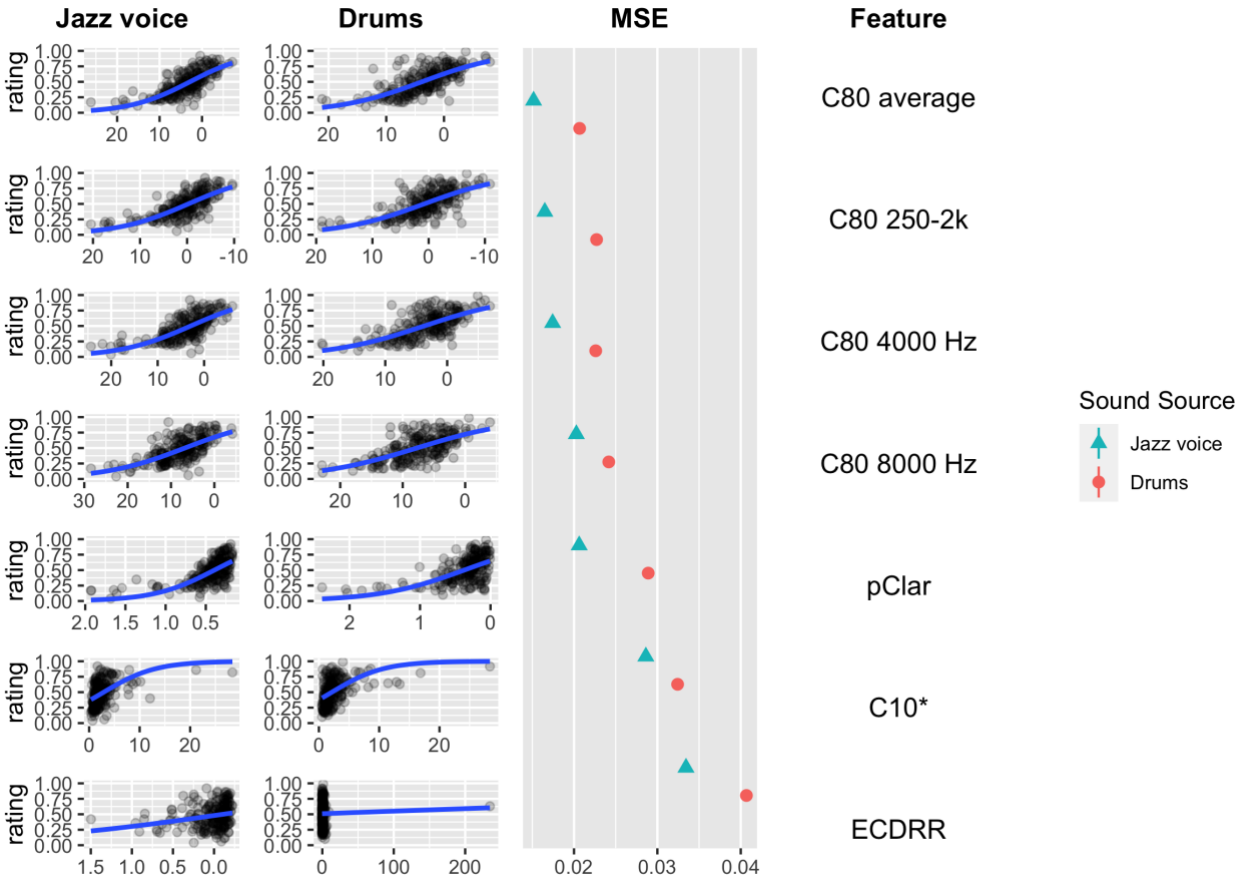


Figure 23 Candidate models for *distance*

As discussed earlier, the *ECDRR* calculates a DRR in a two-step process, by first estimating the direction of arrival of the direct sound, and then attempting to remove, or cancel, energy arriving from that direction. In situations such as this outlying IR, where the left and right channels are nearly identical, the cancellation operation is extremely effective. The vast majority of the IR's energy can be cancelled, leaving very little residual energy. This results in a low estimate of the "reverberant" energy, and hence a large *ECDRR*.

Taken to an extreme, this characteristic of the model implies that all purely monophonic IRs, where the left and right channels are identical, will have infinitely large *ECDRRs*. That is, binaural signal differences are required for *ECDRR* variation. This fact is somewhat at odds with psychoacoustic findings, which indicate that differences in DRR can be perceived even in the absence of binaural cues (Larsen et al., 2008). Since the *ECDRR* cannot predict DRR differences in monaural signals, it will likely also fail to predict differences in source distance if the cues at play are monaural in nature.

In summary, this outlier suggests a shortcoming in the *ECDRR* as a distance predictor. As the *ECDRR* is insensitive to monaural DRR cues, it can be expected to give poor predictions when distance is cued monaurally. This appears to have been the case with this stimulus, which had a relatively high distance rating (0.63) despite having nearly identical signals in the left and right channels.

4.4.1.1.2 C_{10}'

Additionally, a handful of IRs have unusually large C_{10}' values. In the jazz voice column, two IRs are above 20, while in the drums column, one is above 20. An investigation into these outliers determined that both were associated with IRs with long attack times, that is, IRs whose peak occurred long after its onset. Such IRs might arise from geometric configurations in which there was no direct path between the measurement loudspeakers and the microphone array: for instance, if the loudspeakers were positioned in an orchestra pit and the microphones were in the audience area.

The C_{10}' feature, as explained earlier, includes an energy ratio between a 10 ms window beginning at the IR's onset, and a longer window containing the remainder of the IR. In cases where the IR's peak occurs more than about 10 ms after its onset, relatively little energy will fall in the first window, and a comparatively large amount will fall in the second. This can result in a very large C_{10}'.

4.4.1.1.3 Effects of outliers

To investigate the effect of these outliers on the results, the data were reanalyzed with outliers removed. This new data set is shown in Figure 24. In this second analysis, in which all $C_{10}{}'$ values greater than 6 (28 data points in total), and the single *ECDRR* value of 234 were excluded, the MSEs of the models changed slightly, but the rankings of the models did not. Outlying values of the $C_{10}{}'$ and *ECDRR* features, then, did not appear to have an undue influence on the results.

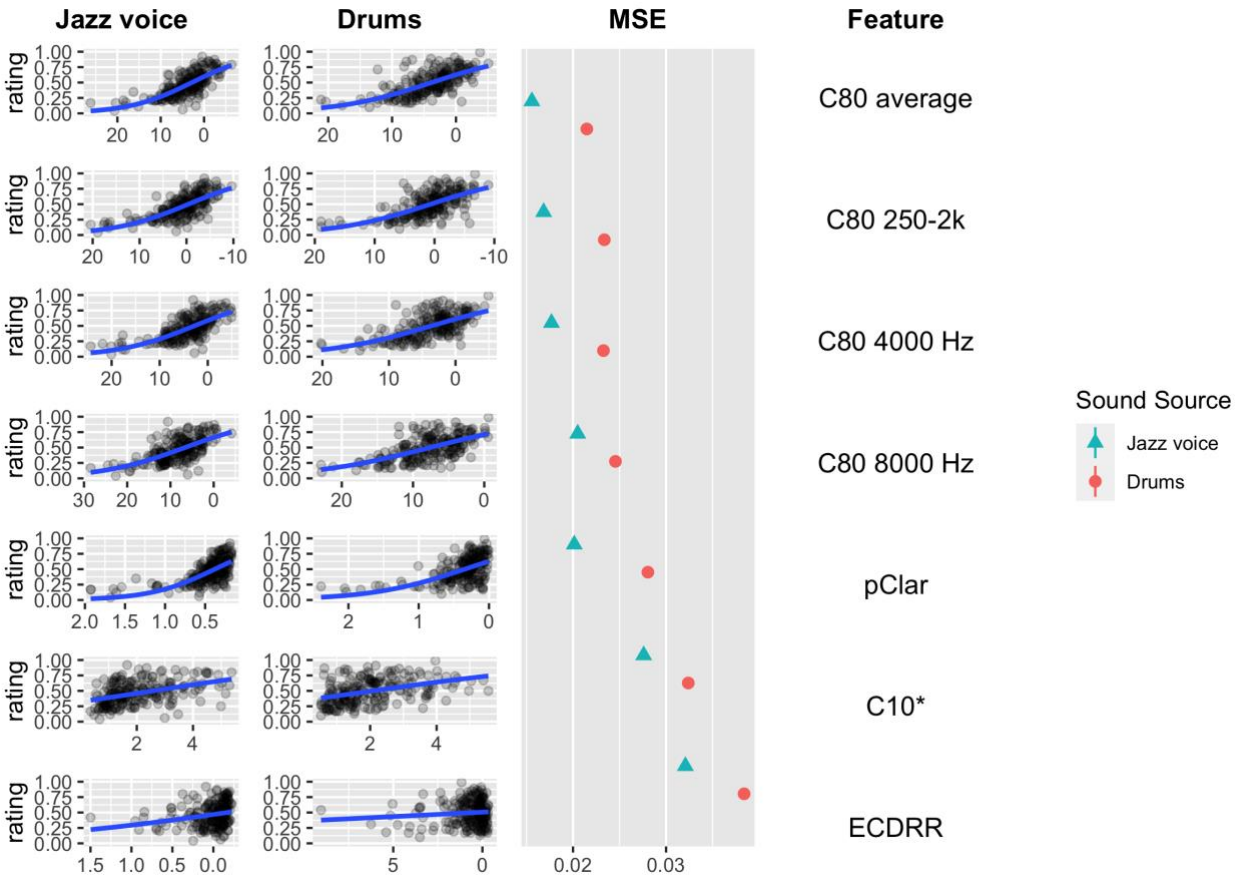


Figure 24 Candidate models for *distance*, outliers removed

4.4.2 Brightness

Unlike the distance attribute, for brightness, the best fitting model depended on the sound source. For drums, the *spectral slope (IR)* had the lowest MSE, while for jazz voice the *spectral decrease* had the lowest MSE. In Figure 25, the *spectral slope (IR)* appears at the top as it had the lowest value when the MSEs from both sound sources were averaged.

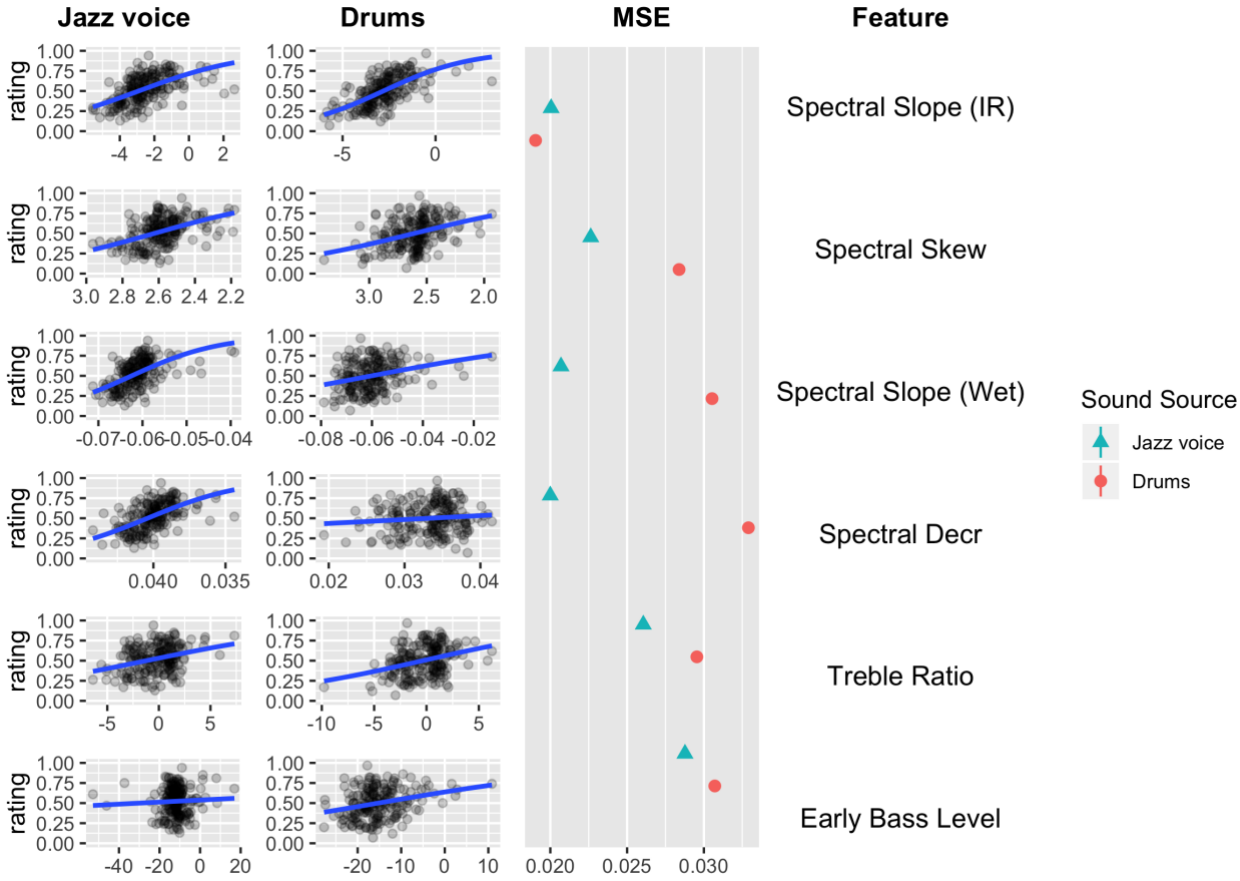


Figure 25 Candidate models for *brightness*

4.4.3 Significance tests for top-scoring features

An important question about these results concerns whether the differences between the top scoring models (C_{80} *average*, *spectral slope (IR)* and *spectral decrease*) and the other models is "significant". Or, in other words, whether the differences between the top models and the other models is larger than the sampling error that would be expected on these differences.

This question can be answered by building confidence intervals on the differences between the top models and the remaining models, and then checking whether these confidence intervals include zero. If the interval does not include zero, the performance difference between the models is significant.

4.4.3.1 Distance

Figure 26 shows 95% confidence intervals for the differences between the C_{80} *average* and the other six models. The intervals were generated from 10,000 bootstrap samples of model differences, using the bootstrap percentile-adjusted (BCa) method (Canty & Ripley, 2021; Davison & Hinkley, 1997). For the drums source, the C_{80} *average* MSE is not significantly different from that of the C_{80} *250-2k* or the C_{80} *4000*. The difference from the other models is significant. The results for the jazz voice source are nearly identical, except that in this case C_{80} *average* outperforms five other models instead of four.

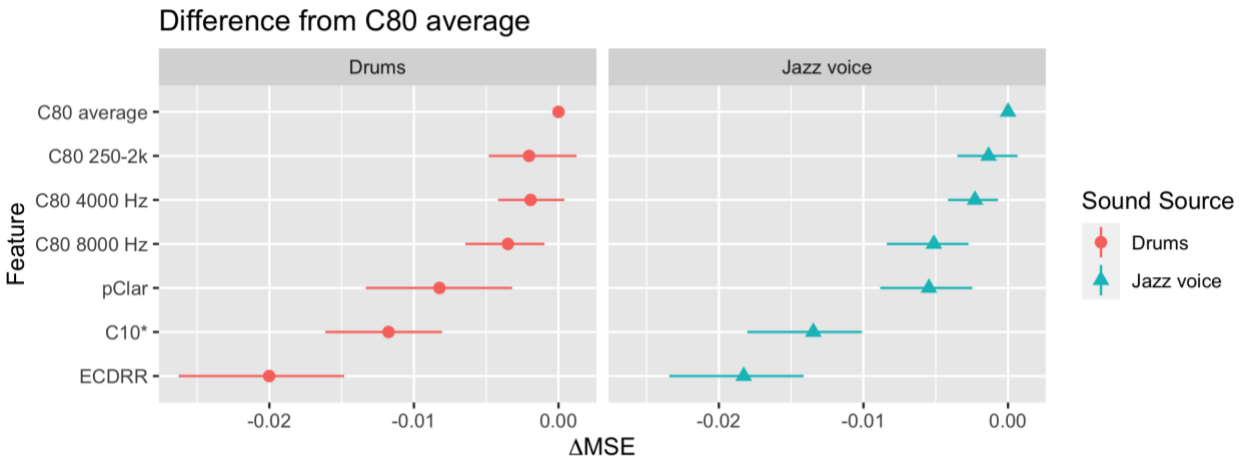


Figure 26 MSE difference from *C80 average*

4.4.3.2 Brightness

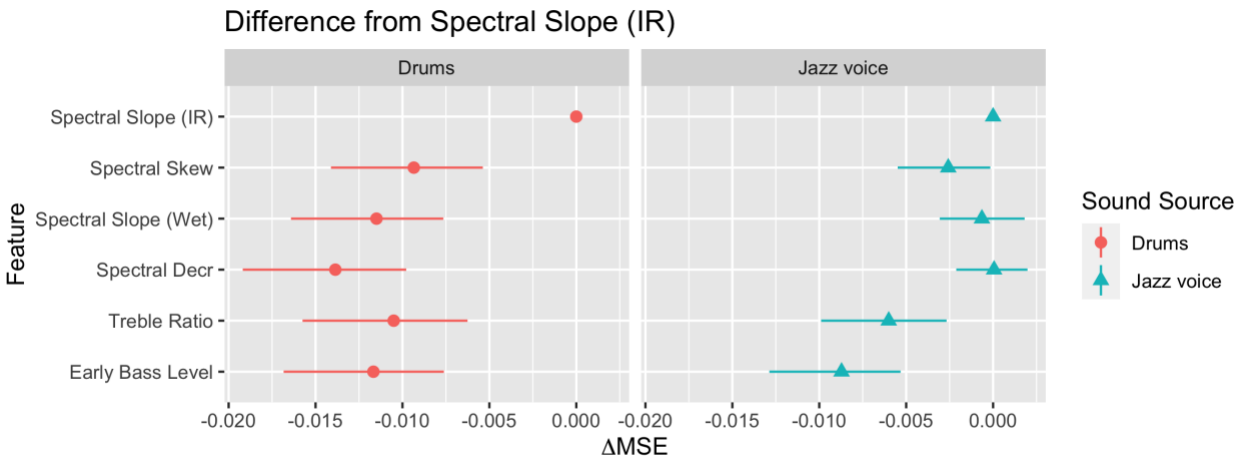


Figure 27 MSE difference from *Spectral Slope (IR)*

For the brightness data (Figure 27), the results depend more strongly on the sound source. For drums, *spectral slope (IR)* has a significantly lower MSE than all other models. In the case of jazz voice, however, two models show comparable performance to the *spectral slope (IR)*, namely the *spectral slope (wet)* and *spectral decrease*. Statistically speaking, on the jazz voice source, the *spectral slope (IR)*, *spectral slope (wet)*, and *spectral decrease* models all have equivalent performance.

4.5 Discussion

This section considers the experimental results in relation to the hypotheses outlined earlier.

4.5.1 Source Distance

The introduction to this chapter suggested that the *ECDRR*, being a wet feature explicitly designed to predict source distance, would have the best fit to our data, followed by the C_{10}', and then followed by the features designed to predict perceived clarity rather than distance: *pClar*, C_{80} *average* and C_{80} *250-2k*. These hypotheses were not confirmed. Surprisingly, rather than performing worst, the C_{80} *average* and C_{80} *250-2k* outperformed almost all other models, and, indeed, the *ECDRR* performed worst.

The high MSEs of the *ECDRR* and C_{10}' features might be partly explained by a poor ability of these features to predict distances in large enclosures and with long reverberation times. The publications introducing the *ECDRR* and Bronkhorst's distance model, from which the C_{10}' is derived, evaluated these features in spaces that were generally smaller and less reverberant than those used in the present study. Lu and Cooke (2010) tested the *ECDRR* in a 9x6x4 m classroom and in a virtual space with a reverberation time $T$of 0.7 seconds; Bronkhorst and Houtgast tested their measure in environments with T ranging from 0.1 to 2 s. The IRs used in the present study, by contrast, were considerably more reverberant, with a median T_{30} of 2.3 s. This means

that well over half of the stimuli used in this experiment were more reverberant than those used to validate the aforementioned models. It is possible that the excellent performance of these models documented in smaller spaces does not generalize well to larger ones.[33]

A second explanation for the high MSE of the $C_{10}{}'$ model may relate to the differences between Bronkhorst and Houtgast's original distance model and the simplified version of it used here. As stated earlier, Bronkhorst's original model also incorporated the room's reverberation radius; this data was not leveraged by the $C_{10}{}'$. A distance model closer to Bronkhorst's original conception could have been studied in this investigation with additional effort, but such a model would have defeated one of the aims of the present work, which was to predict reverberation attributes using only information present in the IR and the wet signal, and without the additional metadata required to calculate the reverberation radius (i.e. the room's volume and the sound source's directivity).

4.5.2 Brightness

Concerning brightness, the introduction section divided the candidate models into three groups: features based on the wet signal (*spectral slope (wet)*, *skew* and *decrease*), a novel IR-based feature (*spectral slope (IR*)), and two IR-based features introduced by Soulodre and Bradley (1995) (*early bass level* and *late treble ratio*). The wet features were expected to perform best, followed by the *spectral slope (IR)*, followed by Soulodre and Bradley's features. These expectations were partially met, as the wet features did indeed have lower MSEs than Soulodre and Bradley's. The *spectral slope (IR)*, however, exceeded expectations by performing either on par with the wet features, or better than them, depending on the sound source used.

The superior performance of *spectral slope (IR)* relative to the wet features is difficult to explain. One possibility relates to temporal window used by the *spectral slope (IR)*.

[33] The *ECDRR*'s poor performance here may also be related to its origin as an analysis tool for dummy head microphone signals, rather than for signals from widely spaced microphone pairs. The relevant differences between these two signal types are discussed in appendix B, specifically in the section on the *pRev*, *pClar, pASW* and *pLEV* features.

This feature examined only the IR's initial 300 ms (beginning at its onset) and ignored the remainder. The wet features, by contrast, made no explicit distinctions between early- and later-arriving sound and were influenced by the entire IR. It is possible that the attribute of brightness, in the specific context of reverberation, is dominated by information near the beginning of the IR (e.g. early reflections arriving in the first few hundred milliseconds) and less influenced by later sound. If this were the case, future productive work might involve refining this temporal window to see if longer or shorter durations lead to better brightness predictions.

4.6 Conclusions

The aim of this chapter was to develop predictive models of brightness and source distance, two attributes that, judging by the investigation in chapter three, appear to be salient sources of perceptual variation within natural IR libraries. A listening test was conducted to collect ratings on these attributes. The test yielded brightness ratings on 406 different stimuli and distance ratings on 476 stimuli.

Thirteen candidate models for predicting these ratings were examined. Most models were drawn directly from the existing literature on psychoacoustics and room acoustic perception, although one, the *spectral slope (IR)*, was novel and introduced in this chapter. When the models were compared with respect to their ability to predict distance, those based on the traditional *C80* feature fared best. The two most effective specific models were the C_{80} *250-2k* and the C_{80} *average*, which averaged the 125-8000 Hz octave bands.

For predicting brightness, the optimal model depended on the sound source used. For drums, the newly proposed *spectral slope (IR)* appeared most effective, while for jazz voice, no statistical differences were found between the *spectral slope (IR)* and two features based on the wet signal, *spectral slope (wet)* and *spectral decrease*.

Overall, in the cases of both brightness and distance, it was somewhat surprising that such accurate predictions could be achieved by models that examined only the IR, as opposed to models that operated on the complete wet signal.

Finally, the previous chapter raised questions about the independence of the attributes of source distance and perceived clarity. To help determine whether the two attributes were distinct, that is, whether it were possible for natural reverbs to sound simultaneously both *close* and *unclear* or both *far* and *clear*, the previous chapter suggested searching for a signal feature that could predict one attribute significantly better than the other. Perceived clarity is thought to be well predicted by the C_{80}. If some other feature were found, then, that could predict source distance better than C_{80}, this would provide evidence to support the notion that clarity and distance were independent. This chapter failed to identify such a feature. Despite examining several

promising distance models from the literature, no objective signal feature could be found that predicted source distance better than the C_{80}. As such, this chapter provides no evidence to support the notion that the attributes of source distance and perceived clarity are perceptually distinct, and capable of varying independently of one another.

4.6.1 Final model selection

The rationale for this investigation was to identify predictive models that could be used to enable searches along dimensions of brightness and distance in a perceptual search interface for IRs. In this context, it is fortunate that IR-based models proved so apt, as these models are more computationally convenient. IR-based models depend only on the IR data in the library, and as such, can be accurately computed offline. Wet signal-based models, by contrast, depend on both the IR and the stem or live signal chosen by a mixing engineer. This stem or live signal is generally not known in advance of the mixing session. To use these features in a search interface, then, would either require expensive computation during the mixing session, or a computational shortcut that might sacrifice accuracy (such as computing features in advance on a set of representative sound sources which would not necessarily include the specific source used in production).

The fact that some models can be computed offline can help us select the two most appropriate models for search interface applications.

For the brightness attribute, the better-performing models were the *spectral slope (IR)*, *spectral slope (wet)* and the *spectral decrease*. of these, the *spectral slope (IR)* seems the wisest choice, as it is both computationally convenient, and gave significantly better results with the drums source.

For the distance attribute, the best performing models were the C_{80} *250-2000* and the C_{80} *average*. Of these, the C_{80} *average* seems the wiser choice, as it is expected to be more robust to variations in sound source frequency content. The C_{80} *250-2000*, by contrast, might perform poorly on sound sources lacking significant energy at mid-frequencies.

In the next chapter, these two models will be used to create perceptually informative visual representations of IRs. These visual representations will be designed to facilitate IR browsing within large collections.

5 EVALUATION OF ROOM IMPULSE RESPONSE VISUALIZATIONS

5.1 Introduction

Earlier chapters of this work attempted to develop a perceptual model of natural reverberation. More precisely, they sought to describe the dimensions of perceptual variation of reverberation, and to develop signal processing techniques to predict values on these dimensions. Chapter two reviewed existing work in these areas, focusing on the model presented in ISO 3382 and its four dimensions of perceived reverberance, perceived clarity, ASW and LEV (International Organization for Standardization, 2009). Chapter 3 explored additional dimensions beyond these four, and proposed two further dimensions named source distance and brightness. Chapter 4 investigated distance and brightness and derived signal processing techniques to predict them.

This chapter will synthesize these earlier findings into a revised model of reverberation. This revised model will then be used to create visualizations of impulse responses. These visualizations, or *glyphs*, will be designed to support rapid searches within large IR libraries for items with particular perceptual characteristics. That is, the glyphs will be designed to facilitate the kinds of searches carried out by mixing engineers: searches where reverberation with a particular character is sought, and where the engineer wishes to find the impulse response in the library that most closely matches this desired target.

Due to humans' capacity for rapid visual search, glyph visualization can be a valuable tool for searching within datasets (Ware, 2021). The performance of such glyph-based search interfaces depends strongly, however, on whether the visual characteristics of the glyphs are well matched to the capabilities of the human visual system (Wolfe et al., 1989). Thus, understanding human vision is critical for effective glyph design. This chapter will begin with a short review of relevant research on human vision, along with results from existing work on glyph design. Inspired by these results, a novel glyph design for room impulse responses will be proposed, and this novel design will be evaluated in a controlled experiment. Results will be presented, and the implications for large impulse-response library search interfaces will be discussed.

5.2 Glyph design and human vision

In the context of information visualization, a glyph is visual representation of a multidimensional data point. It consists of an icon whose visual features convey the values of the underlying datum. Glyph-based displays are particularly useful for revealing associations between sets of two or three variables, or for identifying data points with particular values on a small combination of variables. In a reverberation context, this means that a well-designed glyph should make it easy to visually identify impulse responses with, say, low reverberance, but high width and brightness.

This section reviews some research on glyph design, centering on a popular format known as a "star" or "whisker" glyph. Star glyphs have the appealing quality of being domain-agnostic: they work equally well in any application domain. Following this discussion of domain-agnostic star glyphs, two strategies will be presented that are often used in the creation of more domain-specific glyphs. These strategies are "natural mappings" and "visual channel separation". Natural mappings attempt to accommodate limitations in human short-term memory, while visual channel separation uses knowledge about human vision to facilitate visual search.

5.2.1 Star glyphs

Glyph design is a well-studied topic with a history dating back to at least the 1950's (Anderson, 1957; Fuchs et al., 2017). Throughout the course of the last half century, however, one of the most enduring designs has been the star or whisker glyph. In such a design, each data attribute is represented by a "ray" emanating from a central point, with the ray's length denoting the attribute's value. Many elaborations on this basic design exist, such as contour lines to connect the ends of the rays (Chambers et al., 1983), or tick marks to indicate reference values. Experiments have shown, however, that "bare" plots, without reference values, are surprisingly effective at supporting data similarity judgements, and that most common elaborations are either unhelpful or counter-productive (Fuchs et al., 2014). An example of a star glyph design used to visualize a set of three-dimensional data points is given in Figure 28.

Despite their long history and widespread popularity, however, star glyphs exhibit several disadvantages compared with glyphs constructed using more modern design principles. For one, the abstract nature of star glyphs can make them difficult to interpret: the glyphs typically provide no visual clues as to the relationships between rays and the data variables the rays represent. For another, the graphical properties of star glyphs can complicate visual searches, as explained below (section 5.2.3.3). More recent writings have suggested that superior designs can be created by following the principles of *natural mappings* and *visual channel separation* (e.g., Borgo et al., 2012; Karve & Gleicher, 2007).

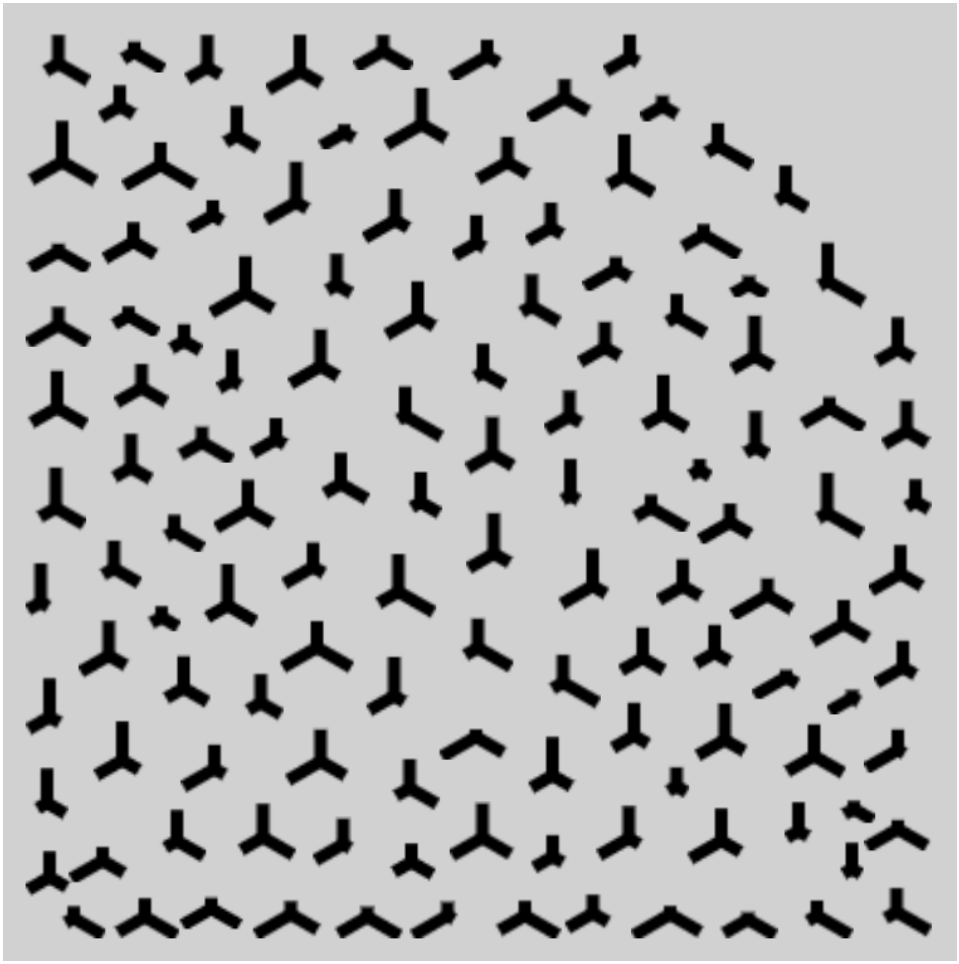


Figure 28 Three-dimensional data points visualized with star glyphs

5.2.2 Natural mappings

The term natural mapping emerged in the field of user interface design in the 1980s (Norman, 1988). Here, a "mapping" describes a relationship between an action on a device and the outcome resulting from that action. If the outcome can be intuitively predicted from the action, the mapping is said to be "natural"; if the outcome cannot be intuitively predicted, the mapping is "unnatural".

Simple examples of natural and unnatural mappings can be found in the design of bicycles. For instance, a natural mapping exists in a bicycle's steering mechanism. While the bicycle is moving, the action of rotating the handlebars is mapped to the outcome of the bicycle turning. This mapping is natural because the outcome can be intuitively predicted from the action. By quickly examining a bicycle, even a non-cyclist will notice that the handlebars are connected to the front wheel, and that rotating one will rotate the other in the same direction, causing a moving bicycle to turn to the left or right. Learning the handlebar-to-turn-direction mapping requires little thought and no explanation; for this reason it is natural.

By contrast, a bicycle's gear shifter exhibits a less natural mapping. Without prior experience, it's generally difficult to know in which direction a shifter should be pushed to ride easily up a hill. Learning the mapping between shifter position and pedal effort requires either trial and error or a careful examination of the bicycle's derailleur. Understanding the mapping requires cognitive effort, and this makes it less natural.

Natural mappings are desirable in user interfaces because they lessen the burden on a user's working memory. Whereas unnatural mappings must be held in memory, natural mappings are obvious enough that they don't need to be remembered, and, as such, leave more working memory available for other tasks related to the user's objective. Likely due to these memory advantages, as well as to their ease of learning, natural mappings have been shown to be associated with concrete speed gains on certain experimental tasks (Froehlich et al., 2006).

5.2.2.1 Natural mappings in glyph design

In the domain of glyph design, natural mappings refer more specifically to easy-to-understand relationships between data attributes and a glyph's visual features. As an example, consider the challenge of creating a glyph design to visualize a hypothetical dataset consisting of three weather-related variables: *air temperature, wind direction* and *wind speed*. In each of these three cases, strong cultural conventions exist regarding the attribute's visual representation. In meteorological contexts, air temperature is often mapped to colour. In cartography, cardinal directions are mapped onto orientations, with north pointing upward. In more general contexts, magnitude is often mapped to icon size. Taken together, these conventions suggest that a relatively intuitive mapping from these three attributes onto visual glyph features might involve mapping *air temperature* onto colour, *wind direction* onto orientation, and *wind speed* onto icon size. An example of a set of glyphs constructed with this mapping is shown in Figure 29. While this mapping isn't perfectly unambiguous, it's likely to be relatively easily understood because it draws on familiar conventions. If asked to select the icons corresponding to strong, cold, northernly winds, most users would likely select large, dark icons pointing upward. Because these mappings can be understood without reference to a legend or description, they are relatively natural.

By contrast, an "unnatural" mapping of similar data can be seen in the star glyph example in Figure 28. Like Figure 29, this figure also depicts a set of three-dimensional data points and could, in principle, also be interpreted as a visualization of the three weather-related variables. In this latter case, though, identifying the icons corresponding to strong, cold, northernly winds would be impossible without the help of a legend. For this reason, the mapping it uses is unnatural.

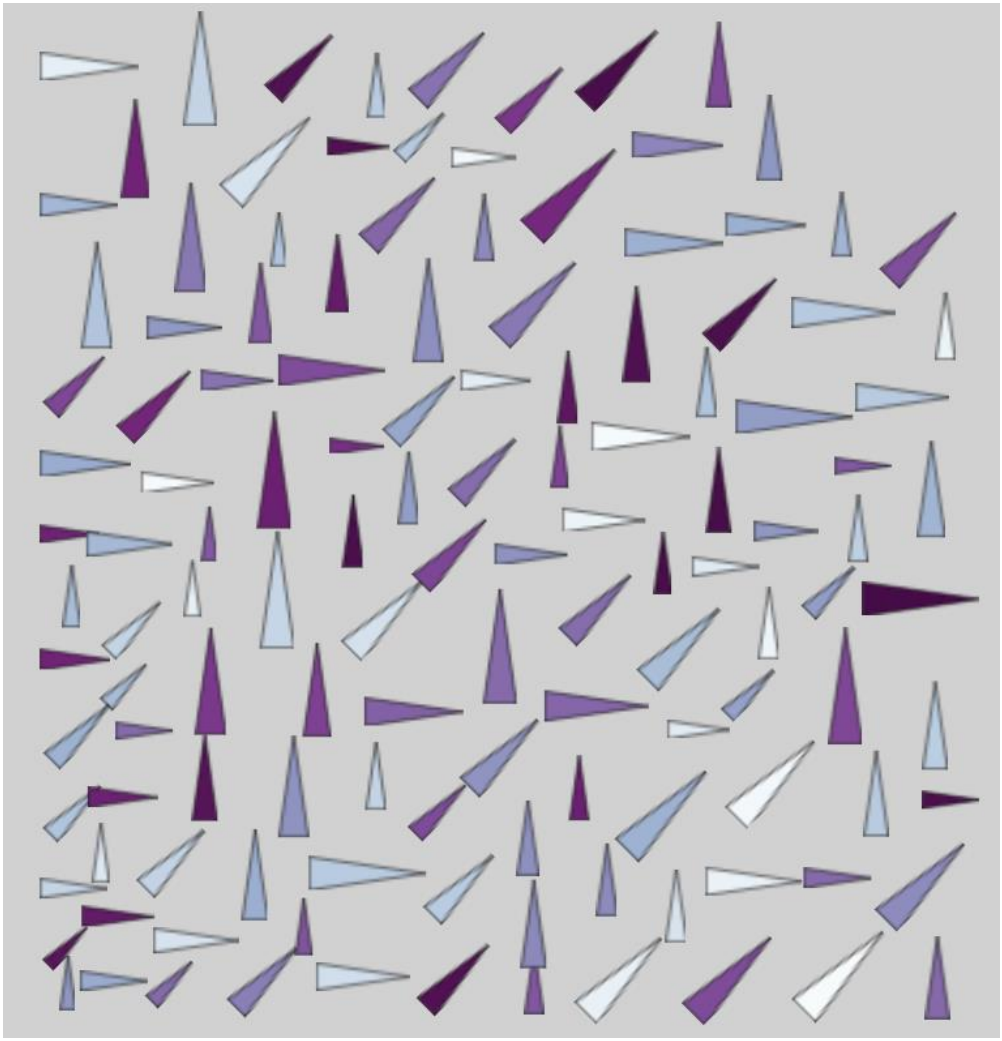


Figure 29 Three-dimensional data glyphs exhibiting natural mappings for *wind speed, wind direction* and *air temperature*

As with natural mappings in general user interfaces, natural mappings are desirable in glyph design because they lessen the burden on a user's working memory. When mappings between visual features and data attributes are intuitive enough that they don't need to be held in working memory, more memory is available for other tasks. In addition to their memory efficiency, under certain conditions, natural mappings have also been shown to be associated with faster and more accurate performance on visual search tasks (McDougall et al., 2000).

5.2.3 Visual channel separation

After natural mappings, the second design principle relevant to the construction of our impulse response glyphs is that of visual channel separation. This principle builds on the conceptual foundations of the *visual channel*, the *visual channel map,* and the relationship between *visual channel maps* and visual search speed. These ideas will be introduced next.

5.2.3.1 Visual channels and channel maps

In vision research, *channels* refer to types of visual information that are processed in parallel by the early visual system (Ware, 2021; Wolfe et al., 1989). Basic channels include luminance, red/green colour differences, blue/yellow colour differences, motion, and elements of form and texture. The form and texture channels can be further divided into sub-channels that track particular spatial frequencies and object orientations.

Channel theory contends that the neural signals output by the light-sensitive cells in the retina are passed on to an enormous network of neurons in the early visual system. Portions of this network operate in parallel to produce spatial "maps" of the visual field, each tuned to a particular channel. Thus, luminance maps highlight the locations of bright and dark regions in the visual field; colour difference maps highlight locations with green but little red or blue but little yellow. Orientation maps, activated by right

or left slanted edges, highlight the locations of right or left slanted objects. Spatial frequency maps highlight the locations of large or small objects.

An illustrative example of channel maps is given in Figure 30. Here, the first panel shows a scene consisting of bars, circles and triangles; the following three panels show how this scene would be represented in three different channel maps. The first map is tuned to red/green differences and shows three areas of high activation corresponding to red objects. The second is tuned to low luminance and shows activity in three different regions, corresponding to the black objects. The third map is tuned to right-slanted object and shows two areas of strong activation and two areas of weaker activation. The areas of strong activation correspond to right-slanted bars, and the areas of weak activation correspond to right-slanted triangles.

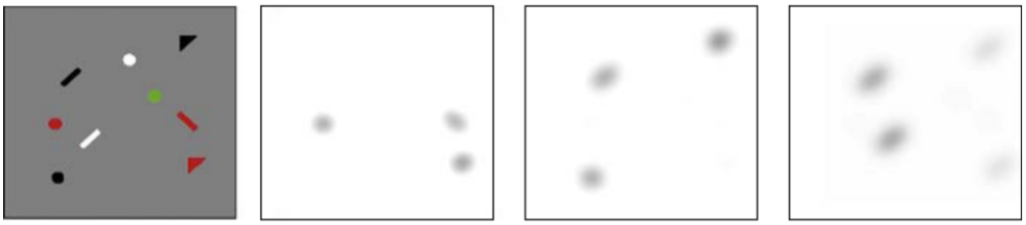


Figure 30 A scene and its representation on three visual channels[34]

5.2.3.2 Channel maps and visual search tasks

With each glance at the world, the early visual system produces a multitude of channel maps. These maps are created automatically, effortlessly and nearly-instantaneously. They generally remain below the level of conscious awareness, but the information in them is used to guide eye movements and direct visual focus. Channel maps are exploited heavily during visual search tasks.

[34] Reprinted from Ware (2021) with permission from Elsevier.

One such search task might involve counting the number of red objects in Figure 30. If asked to carry out this task, the visual system will first take a quick glance at the page. From this glance, channel maps will be generated. The conscious mind would then consult the particular map(s) relevant to the search task at hand. In a search for red objects, the "red" map shown in the second panel will be used. Guided by this map, fractions of a second later, this eye will rotate slightly to focus on each of the three positions highlighted in the map. While receiving focus, the objects can be counted. Each of these "rotate and focus" actions, also known as a "fixation", is relatively slow, requiring hundreds of milliseconds. As this particular search task would require about three fixations, it can be accomplished quickly. The red objects can likely be counted in less than a second.

This task is speedy, however, only because the target items (red objects) are highlighted on a channel map. If the targets of a visual search task happen not to stand out distinctly on any channel map, the search will be much slower. One example of a slower search task might involve counting the number of times the letter "u" appears in the previous sentence. Since no visual channel is tuned to the specific shape of the letter "u", channel maps cannot be used to efficiently guide the search. As a result, although the number of targets present is the same (three), many more visual fixations will be required as the page is scanned. The search will take much longer to carry out, perhaps on the order of several seconds. Channel maps, then, can be used to guide visual searches, and visual searches will be faster if they're able to exploit channel map information.

5.2.3.3 Visual channel separation

Enabling visual search is often an implicit goal of glyph design. As such, it follows that aligning data attributes with visual channels is considered good design practice (Ware, 2021, p. 150). A related, but more subtle design principle is that of *visual channel separation*. Visual channel separation involves ensuring not only that each data attribute is mapped onto one or more visual channels, but, further, that these sets of visual channels are maximally different from one another.

Visual channel separation is most easily illustrated by example. Consider again the contrasting glyph designs in Figure 28 and Figure 29. As before, imagine that both

figures are being used to visualize the same weather-related data. Further, imagine that the *air temperature* variable, mapped to colour in Figure 29, is mapped to the lower-right ray of the star glyphs in Figure 28.

Now consider the degree to which each glyph design supports one particular search task. The task is to locate all icons that represent high air temperatures. With the star glyphs in Figure 28 this means locating all icons with long lower-right rays; in Figure 29 this means locating all icons that are bright in luminance (or close to white in colour). In Figure 29, this task is easy. When searching for the white icons, the targets almost seem to pop out of the page. In Figure 28, by contrast, the search is harder. The image seems "busier" and finding the target icons requires more mental and visual effort.

The difference in ease between these two searches is explained by visual channel separation. In Figure 29 the three data attributes of *temperature*, *direction* and *speed* are mapped to separate visual channels (*colour*, *size*, *orientation*). In Figure 28, by contrast, the three data attributes are all mapped to the same channel (*size*). The lack of channel separation has the effect of "cluttering" the channel maps used in the search. When searching among the star glyphs, a size map is consulted, but this map is excited not only by the target objects (long lower-left rays) but also by distracting objects (long rays in other directions). As a result, the size map contains many false positives, and it incorrectly directs the eye to irrelevant locations. As each fixation on a different location takes time, these false positives slow the search. In Figure 29, however, a luminance map is used in the search, and this map is excited only by the target objects. It contains no false positives. As a result, the map can direct fixations more efficiently. Fewer fixations are required, and the search is completed more quickly. The separation of data attributes onto distinct visual channels is responsible for the speed increase.

In summary then, two notions from vision science relevant to glyph design are the *visual channel* and the *visual channel map*. Channels correspond to types of visual information that are processed quickly and automatically by early vision. Channels populate channel maps, which contain the spatial locations of visual features within the field of view, and which are used to guide visual focus (i.e., to plan fixations). Glyph designs work best when data attributes are mapped onto visual channels, and, further, when the channels used by each data attribute are well separated from one another. The

strategy of mapping each data attribute onto a distinct set of visual channels is known as *visual channel separation.*

Having now introduced the design principles of natural mappings and visual channel separation, the remainder of this chapter will apply these ideas to the creation of a novel glyph design for room impulse responses. The novel design will aim to facilitate visual searches along the key perceptual dimensions of room reverberation. In a controlled experiment, this novel design will then be contrasted with a star glyph design that exhibits lower channel separation and less natural mappings. In the following section, the specific hypotheses of the experiment will be presented, and the experiment's methods will be outlined.

5.3 Experimental hypotheses

The experiment described later in this chapter aims to test three specific hypotheses about a novel glyph design for room impulse responses. The novel design will be created using the principles of natural mappings and visual channel separation, as outlined above.

5.3.1 Hypothesis 1: the novel glyphs will be easy to interpret

The novel glyphs used in the experiment will be designed to be easily interpretable by sound engineers. That is, subjects in the experiment are expected to have little trouble understanding the mappings between the visual features of the glyphs and the auditory attributes of reverberation that these visual features represent. In other words, the mappings will be natural. Mapping naturalness will be explicitly tested in the first part of the experiment.

5.3.2 Hypothesis 2: both novel and star glyphs will enable better-than-chance performance in an IR search task

The novel glyph design will be compared with a more generic star glyph design in an IR search task. Both designs, however, will be visualizing data using the same perceptual model of reverberation. Since this model, built up in earlier chapters, is assumed to be a useful approximation of the perceptual space of room impulse responses, almost any visualization of its predictions should be at least moderately successful in supporting search tasks. Therefore, since a useful perceptual model is being employed in both cases, both glyph designs should allow users to locate target IRs faster than would be predicted by chance.

5.3.3 Hypothesis 3: in an IR search task, novel glyphs will outperform star glyphs in speed and efficiency

In our search task, to be defined later, the novel glyph design is expected to outperform the star design on several metrics, including the time taken to locate the target and the number of IRs auditioned during the trial.

The novel glyphs should have a speed advantage primarily due to the more natural mappings between auditory and visual features. Although a legend will be available in trials for both glyphs types, the legend is expected to be consulted less frequently in novel glyph trials. Since consulting the legend is a time-consuming activity, these trials should take less time overall.

Additionally, searches with the novel glyph design should be faster due to better visual channel separation. The star glyphs, by comparison, should be associated with more cluttered channel maps, resulting in more fixations and slower searches.

5.4 Methods

Having listed hypotheses, this section will now describe the structure of the experiment designed to test them. First, the perceptual model of reverberation visualized in the glyphs will be presented. The process for creating the novel glyph design will then be explained. Finally, the experiment's methods will outlined, including its subjects, structure, software interface, and auditory and visual materials.

5.4.1 Refining a perceptual model of reverberation

A first step toward designing glyphs for impulse responses involves choosing which perceptual attributes of reverberation the glyphs should visualize. A multitude of perceptual properties of reverberation have been proposed in the literature, as reviewed in chapter 2, and the challenge in this section will be to choose the subset of attributes that will be most useful for searching an impulse response library. This set of attributes will need to be relatively small, since even exquisitely designed glyphs are limited in the number of variables they can effectively convey (Ware, 2021, p. 175). On the other hand, the set of attributes should be large enough to describe a useful portion of the perceptual space of a library. The number of attributes to include in the model is bound by these two considerations.

This section takes a two-stage approach to attribute selection. First, a set of promising candidate attributes are selected from the literature review in Chapter 2 and the investigations in Chapters 3 and 4. These candidates are then listed, along with the objective signal features thought to best predict them. Second, correlations between the associated signal features will be considered. Signal feature correlation information will be used to prune down the initial list of candidate attributes. Promising attributes from the literature review are presented first.

5.4.1.1 Perceptual attributes and signal features from literature review

Considering the literature review, four subjective attributes stand out for inclusion in our model. These are the four present in the ISO 3382 model: perceived reverberance, perceived clarity, apparent source width (ASW), and listener envelopment (LEV).

As discussed in chapter 2, many objective signal features have been shown to correlate well with these attributes. For the purposes of the experiment, however, one feature must be chosen for each. It should be noted that these choices are somewhat arbitrary, since in most cases determining which of several features best predicts the attribute remains a subject of scholarly debate.

For reverberance, the *early decay time* will be used (see section 2.2.2). Additionally, a logarithmic transformation will be applied to make the measure more perceptually uniform. This is justified as the just-noticeable difference of *EDT* is known to grow proportionally with *EDT* magnitude (International Organization for Standardization, 2009). The final reverberance feature will be referred to as the *log EDT*.

For apparent source width and listener envelopment the binaural model features of Schuitman (2013) discussed in section 2.2.2.4 will be used. These are wet features, calculated on the convolution of an IR and a source signal. These features will be referred to as the *sASW* and the *sLEV*.[35]

For clarity the well-known C_{80} will be chosen (section 2.2.2).

5.4.1.2 Perceptual attributes and signal features from chapters 3 and 4

Two additional attributes were identified in Chapters 3 and 4 that seemed effective for characterizing perceptual differences in natural reverberation. In chapter 3 the terms

[35] The versions of the features used in this chapter are slightly different from the versions presented in Chapter 2. The Chapter 2 features *pASW* and *pLEV* are measured in the somewhat arbitrary scale of "model units". By contrast, the versions used in this chapter, *sASW* and *sLEV,* have been transformed by a sigmoid function to have a range between zero and one.

source distance and brightness were chosen to describe these attributes. Later, in chapter 4, these attributes were found to correlate strongly with sigmoid-transformed versions of the C_{80} and the *spectral slope (IR)*. Accordingly, these same two features from Chapter 4 will be chosen as predictors for distance and brightness. The sigmoid-transformed version of these features predicting these attributes are referred to as the C_{80} *distance* and the *IR spectral slope brightness*, respectively.

5.4.1.3 Inter-feature correlations

In total, the two sections above propose six signal features as candidates for inclusion in the model. Before including all six, however, correlations between them will be examined. For a given number of features, choosing the subset with the smallest amount of internal correlation will maximize the amount of information present in the glyphs.

Correlations between these six features are shown in Figure 31. Specifically, the grid illustrates the absolute value of Pearson's r for each feature pair. To estimate the correlations, the same pool of IRs used in the previous chapter's experiment was considered. Wet features *sLEV* and *sASW* were computed by convolving the IRs with the drums sound source (see appendix C), as this sound source was used in the experiment to be described in section 5.4.4.

Examining the matrix, particularly high correlations (> 0.80) are seen to exist between two pairs of features: C_{80} and C_{80} *distance* ($|r| = 0.99$), and *log EDT* and *sLEV* ($|r| = 0.84$).

The near-perfect correlation between C_{80} and C_{80} *distance* is unsurprising given the monotonic relationship between them. C_{80} *distance* is simply a sigmoid transformation of C_{80}. The *log EDT* and *sLEV* correlation was also to be expected given the structure of the latter feature. As discussed in section 2.2.2.4, *sLEV* has two components, one of which is a measure of perceived reverberance similar to *log EDT*.

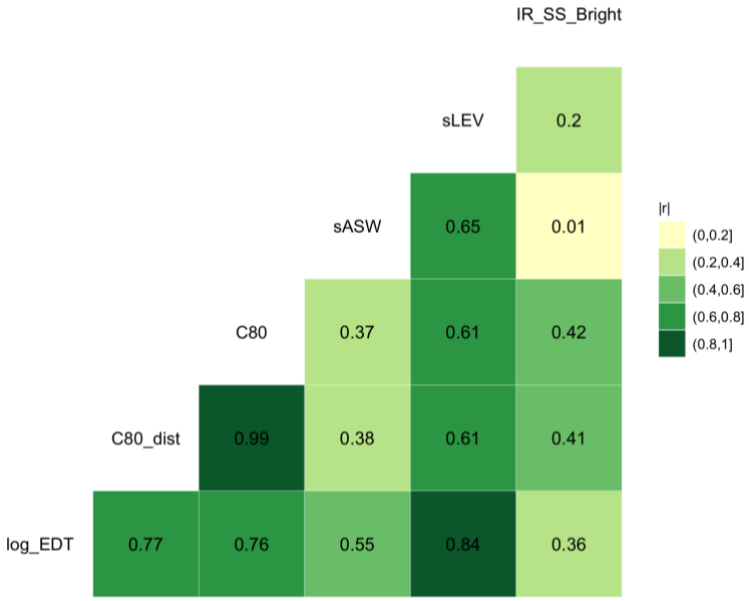


Figure 31 Inter-feature correlations

In both cases, the high correlations between each pair of features suggest that including all four in our model would add visual complexity to the glyphs without a corresponding increase in perceptual information. In order to keep the visual complexity of the glyph design low, only one feature from each pair will be included. Once again, somewhat arbitrarily, C_{80} *distance* will be chosen from the first pair, and *log EDT* from the second.

The final four features chosen to be included in our perceptual model of reverberation are *log EDT*, C_{80} *distance*, *sASW*, and *IR spectral slope brightness*. These four features will be visualized in two glyph designs. In the experiment to follow, the perceptual attributes associated with these features will be described using the names *decay time*, *source distance*, *source width*, and *brightness.*

5.4.2 Creation of the novel glyph design

To create the novel glyph design to be evaluated in the experiment, an iterative process was used. The process was guided by both the author's intuitions and by experimental data. The process consisted of two stages: a design phase and an evaluation phase.

In the design phase, the author proposed a mapping of reverberation attributes onto visual glyph features. A grid of icons was then produced in which each row exhibited monotonic variation in one of the features. An example of such a grid is shown on the right side of Figure 32.

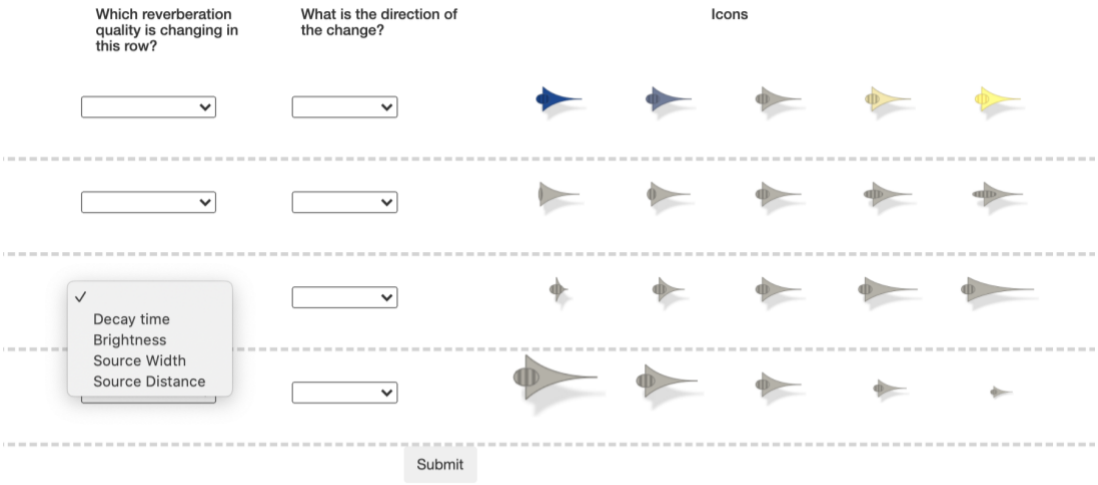


Figure 32 Questionnaire used in glyph design process

In the evaluation phase, the grid was shown to a set of subjects in an informal experiment. The subjects were asked to guess the intended mappings of reverberation attributes onto visual features. As shown in the figure, the four reverberation attributes – decay time, source distance, source width and brightness – were presented in drop-down menus next to each row of icons. For each row, subjects were asked to select the reverberation attribute that they believed the varying visual feature represented.

In the subsequent design phase, subject responses were examined to gauge the naturalness of the tested mappings. Mappings that could be correctly intuited by most subjects were deemed natural and were left alone, while mappings for which subject guesses did not agree with the author's intentions were deemed less natural and were modified. These modifications led to a new set of icons which were then presented to subjects in another evaluation phase.

Three iterations of this design-evaluate cycle led to the mapping shown in Figure 33. In the evaluation of this grid, the mappings for the brightness and decay time attributes (rows one and two) were correctly intuited by all subjects. A minority of subjects, however, reversed the mappings for the source distance and source width attributes (rows three and four). That is, some subjects guessed that the third row represented a change in source width and, the fourth row, a change in source distance, even though the opposite was intended.

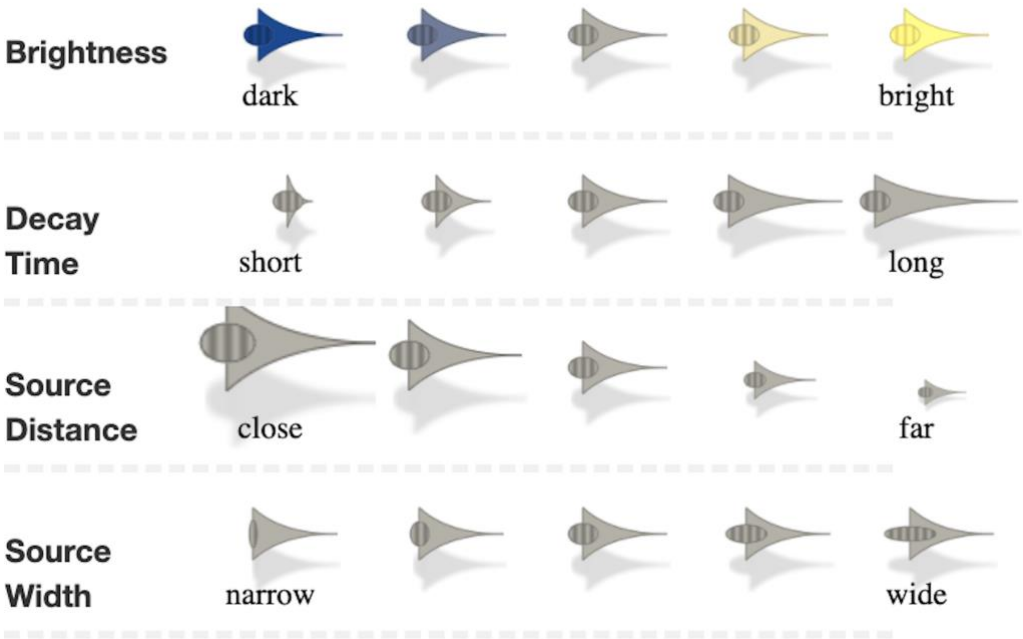


Figure 33 The *novel glyph design*

In a perfect world, the design-evaluation cycle would have been repeated until a perfectly natural mapping was found. Owing to real-world time constraints, however, the process was halted after three iterations in order to proceed with the main experiment.

5.4.2.1 The novel glyph design

The grid shown in Figure 33 is referred to as the novel glyph design. The design consists of four visual features whose values are determined by the four dimensions of the reverberation model. The relationships between visual features, signal features and reverberation attributes are detailed in Table 11.

Table 11 Novel glyph design mappings

Reverberation Attribute	***Signal Feature***	***Visual Feature***
Brightness	*IR spectral slope brightness*	*Colour*
Decay Time	*log EDT*	*Triangle length*
Source Distance	C_{80} *distance*	*Visual distance*
Source Width	*sASW*	*Oval width*

Note that in the novel design, two separate visual cues were used to manipulate visual distance, although only one can be seen in Figure 33. The visible cue is size: icons intended to appear closer to the subject are drawn larger. The second cue is a motion parallax, which was created through an animation effect (Rogers & Graham, 1979).The animation was visible on the webpages where subjects completed the experiment, but cannot be seen on the printed page. To create the motion parallax cue, all icons oscillated slightly along circular paths, with the radius of oscillation inversely

proportional to the intended visual distance. Close icons, such as those on the left of row three, moved in large circles, while distant icons, such as those on the right of row three, moved in smaller circles. In all cases, the radii of the oscillation circles decayed over time, such that all motion had ceased within a few seconds of the subjects first viewing the page. The size and motion parallax cues were meant to reinforce each other, and, together, create a stronger impression of icon distance.

The colour scheme visible in first row, mapped to brightness, follows the Cividis colour map (Nuñez et al., 2018). This colour map is designed to be maximally legible to individuals with colour vision deficiencies. Colour vision deficiencies are thought to affect up to 10% of males (Ware, 2021, p. 98), and, as the sound engineering profession skews male, likely a sizable minority of sound engineers.

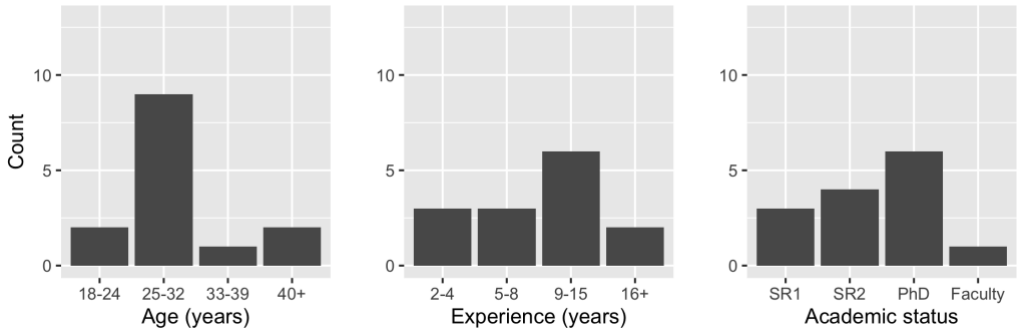


Figure 34 Experimental Subject Demographics

5.4.3 Subjects and recruitment

Experimental subjects were recruited via internal email lists from the Sound Recording Area of the Schulich School of Music. All subjects were graduate students or faculty members in the Sound Recording Area, and all reported at least two years of audio engineering experience. Distributions of subject age, audio experience and academic status are shown in Figure 34. The labels “SR1” and “SR2” refer to the first and second years of a 2-year master’s degree program. Subjects were compensated $20 for their participation. Sixteen subjects took part in the study.

5.4.4 Experiment

To compare the two glyph types, subjects completed an experiment in two parts. The first part was designed to test the naturalness of the mapping in the novel glyph design. The second part was designed to evaluate the ability of the novel glyph design to visually communicate the perceptual character of an IR. This communicative ability was measured indirectly through an IR search task, as detailed below. Following the second part of the experiment, subjects completed a short exit survey to gather basic demographic data and collect subjects' subjective impressions of the novel design.

The experiment was completed over the internet at web pages written by the author.

5.4.4.1 Experiment part one: mapping naturalness in the novel glyph design

The first part of the experiment borrowed heavily from the evaluation phase of glyph design process described in section 5.4.2. In this part, subjects were first presented with written descriptions of the four attributes in the reverberation model. They were then presented with the icon grid and input form shown in Figure 32. As before, subjects were asked to guess the auditory-to-visual mapping in the novel design.

Once subjects completed this task, their answers were immediately reviewed by the experimenter. Subjects were alerted to any cases in which their guesses did not match the intended mappings, for instance, if they guessed that the third row of the figure was meant to depict differences in source width, when in fact it was meant to depict differences source distance. In these cases, the intended mappings that would be used in the second part of the experiment were carefully explained.

Subjects were then directed to a second website, to begin part two.

5.4.4.2 Experiment part two: IR search task performance

The second part of the experiment sought to evaluate the novel glyph design in the context of an IR search task. The task was designed to be vaguely similar to that of choosing a convolution reverb IR during a mixing session: a task familiar to most sound engineers. To establish a baseline performance level against which the novel design could be compared, the performance of a star glyph design on the same task was also measured.

The following sections describe the experiment's software interface, structure, and stimuli.

5.4.4.2.1 Experimental interface

The software interface used in the second part of the experiment is shown in Figure 35 and Figure 36. The first figure shows the interface during a trial in the novel glyph condition, while the second shows a trial in the star glyph condition. In both conditions the interface included a legend showing the current auditory-to-visual mappings at right, and a set of nine icons at left. In the center of the interface was a button labeled "target IR". The target IR was randomly chosen from the set of nine IRs present in the trial. Clicking on the target IR button or any of the icons allowed the associated IR to be auditioned (i.e., convolved with a sound source and sent to the subjects' headphones).

Subjects were instructed to listen to the target, and then try to find the icon that corresponded to the target, all while clicking on as few icons as possible. When the icon representing the target was clicked, the trial ended.

As stated earlier, the experiment was designed to evaluate the ability of each glyph design to visually convey the acoustic character of an IR. Upon hearing the target, it was expected that subjects would visually inspect the nine icons to find the one that most closely corresponded to the target's sound. If the design under test was a "good" visual representation, it was expected that the target icon would be located quickly and with a small number of clicks.

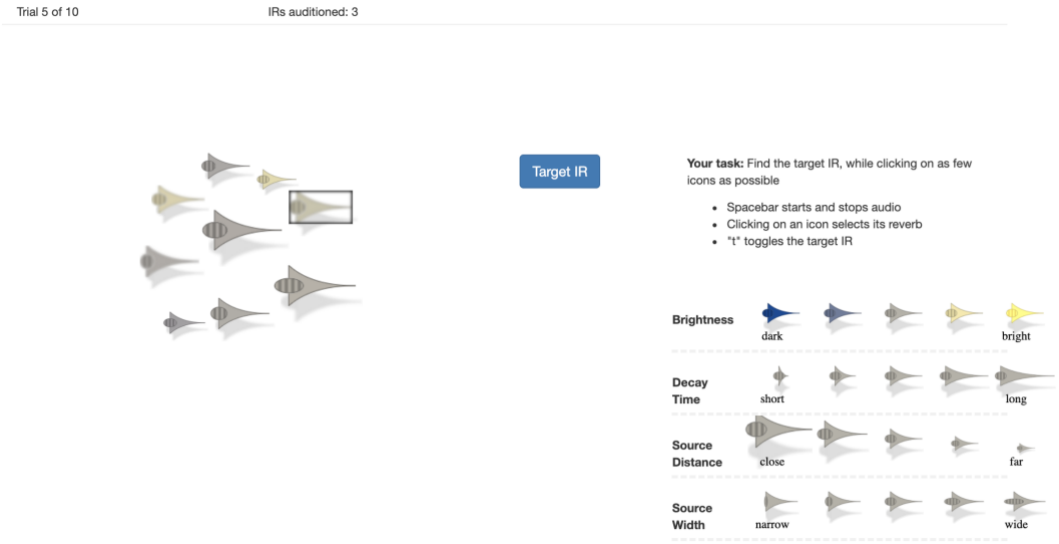


Figure 35 Experimental interface for novel glyph design trials

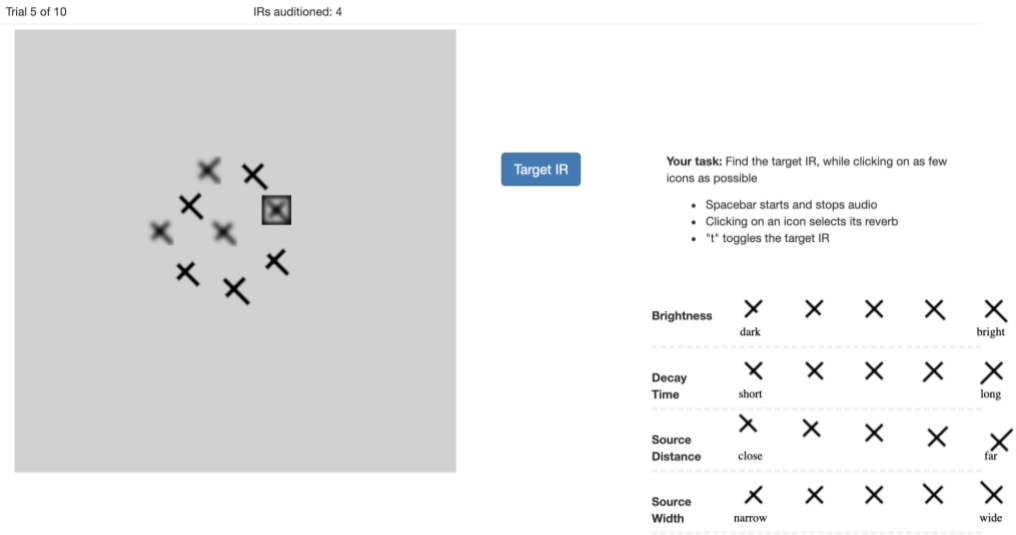


Figure 36 Experimental interface for star glyph design trials

As shown in the figures, subjects were given feedback, through a subtle blur effect, about which icons had already been clicked. They were also given feedback about the number of unique icons already clicked. A black rectangular outline was used to indicate the currently selected IR.

5.4.4.2.2 Experimental structure

The experiment was structured in two blocks, with one block using the novel design and other using the star design. Block order was counterbalanced between subjects, meaning that half saw the novel glyphs followed by the star glyphs, while the other half saw the designs in the reverse order.

Between blocks, subjects were encouraged to take a short break. Each block contained ten trials.

5.4.4.2.3 Audio stimuli: IRs and sound sources

The experiment used the same pool of IRs as the investigation in the previous chapter (see section 4.3.2.1). The nine IRs in each trial were sampled randomly from this pool, subject to the constraint that each trial contained nine unique IRs.

A single sound source file was convolved with the IRs to audition them. In this chapter, the drums recording was used (see Appendix C).

Stimuli were auditioned over headphones. Subjects were able to control playback volume and were instructed to set it to a comfortable level.

5.5 Results

This section examines the results of the experiment and considers whether they lend support to the experimental hypotheses. The first part of the experiment, designed to test hypothesis one, is considered first, followed by the second part of the experiment, designed to test hypotheses two and three.

5.5.1 Hypothesis 1: the novel glyphs will be easy to interpret

The results of the first part of the experiment, the mapping naturalness test, are shown in Figure 37. Rows represent auditory features of reverberation; columns represent visual features of the glyphs. The count in each box indicates the number of subjects who guessed that the given visual feature was mapped onto the given auditory feature. Boxes on the main diagonal represent instances where the subjects' guesses matched the design's intended mappings. Subject guesses and intended mappings matched perfectly for the colour and triangle length features, which were mapped to reverberation brightness and decay time.

The correspondence between guesses and intended mappings was much poorer for the distance and oval width features, however. Here, roughly half of the subjects interpreted these features in the intended way, while half reversed the mapping.

The mappings used in the novel glyph design, then, were not perfectly natural. Some mappings were easy to interpret but others were more ambiguous. Despite the ambiguities in the latter two mappings, though, the process outlined in section 5.4.2 clearly succeeded in creating a design that was at least somewhat intuitive, and that was almost certainly more intuitive than the mapping used in the star glyph design. In the star glyph design, auditory attributes were mapped onto identical-looking rays, and no effort was made to provide visual clues to the mapping's structure.

Figure 37 Mapping naturalness experiment results

5.5.2 IR search task: raw results

Raw results from the second part of the experiment are shown in Figure 38. In each plot, the star glyph condition is shown at left, and the novel glyph condition is shown at right. Dots show the results from individual trials. Means are indicated by horizontal bars.

The top row shows three directly-measured variables related to trial performance. *Unique items auditioned* is the number of unique icons that were clicked before the target was found. This is the variable that subjects were asked to minimize, and on which they were given feedback. *Total items auditioned* is similar but includes repeated clicks. That is, if the icons were numbered from 1 to 9, and a subject clicked icons in the sequence 1-2-1, this trial would be recorded as having 3 *total items* but only 2 *unique items*. *Duration* is the time taken to complete the trial.

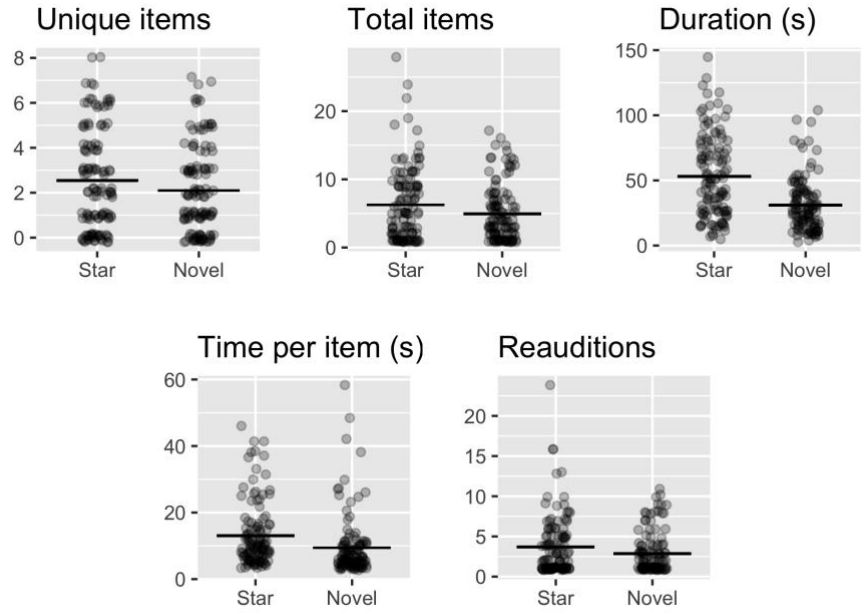


Figure 38 IR search task raw results

The bottom row shows two additional variables derived from the first three. *Time per item* is the average number of seconds between icon clicks and is computed by dividing the trial *duration* by the *total items*. *Reauditions* is the number of clicks icons received following their initial click. Since icons were blurred after their first click, this variable is equivalent to the number blurred icons clicked in the trial. These blurred icons were presumably clicked because subjects were trying to better understand the mappings between visual features and auditory attributes. *Reauditions* is computed as the *total items auditioned* minus the *unique items auditioned.*

Note that of the 16 subjects who participated in the first part of the experiment, only 12 were able to properly complete the second part. Subject attrition was driven by technical failures of unknown causes. In two cases, subjects were unable to connect to the web server hosting the second part of the experiment. In the other two cases, subjects could

access the server but had problems downloading and auditioning the audio material. As a result, data from only 12 subjects is considered in this analysis.

The following sections evaluate the two remaining hypotheses using statistical tests.

5.5.3 Hypothesis 2: both novel and star glyphs will enable better-than-chance performance on an IR search task

Hypothesis two supposed that both glyph types would be at least moderately successful in helping subjects find the target icon. In terms of our experimental variables, better-than-chance performance is equivalent to finding the target IR in fewer clicks than would be expected if the subjects were selecting and clicking icons randomly. Under a null hypothesis of random icon selection, the *unique items auditioned* and *total items auditioned* variables would follow simple theoretical distributions. By comparing these theoretical null distributions with our observed statistics, *p*-values can be computed that describe the probabilities of obtaining our observed results under the hypothesis of random icon selections.

Specifically, under this null hypothesis, *unique items auditioned* would follow a discrete uniform random distribution over the integers 0 to 8. That is, in each trial, if clicking randomly, and if previously clicked icons weren't revisited, between 0 and 8 icons would be clicked before finding the target. The distribution of the mean *unique items auditioned* over 120 trials (12 subjects by 10 trials), could then be derived by convolving this uniform discrete random distribution with itself 120 times (Ruckdeschel et al., 2006; Ruckdeschel & Kohl, 2014). Under this theoretical convolved distribution, the *p*-values of the observed means of *unique items auditioned* (2.55 star design, 2.10 novel design) are both well under 0.001. These results give strong evidence to reject the null hypothesis of random clicking, and support the alternative hypothesis that both glyph designs enabled better-than-chance performance as measured by the mean *unique items auditioned*.

Similarly, under a null hypothesis of random clicking, *total items auditioned* would follow a geometric distribution with a success probability of 1/9; any click would have

a 1/9 chance of finding the target. The distribution of the mean *total items auditioned* over 120 trials, then, would correspond to this geometric distribution convolved with itself 120 times. Under such a null distribution, the observed means of *total items auditioned* (5.84 star, 4.6 novel) again both have *p*-values at or under 0.001. Yet again, these *p*-values provide strong evidence to reject the null hypothesis that the glyphs and the perceptual model were of no use, and suggest that both designs were at least marginally useful.

5.5.4 Hypothesis 3: in IR search tasks, novel glyphs will outperform star glyphs in speed and efficiency

To check for performance differences between the two glyph types, t-tests were performed on the five variables under investigation. Group means and associated *p*-values are shown in Table 12. The sample means of all variables are lower in the novel glyph condition than the star glyph condition. Further, these differences are significant at $\alpha = 0.05$ for all variables except *unique items auditioned*. Significant *p*-values are show in boldface.

Table 12 T-test results on IR search task variables

	$\bar{x}_{star}$	$\bar{x}_{novel}$	$\bar{x}_{star} - \bar{x}_{novel}$	p
Unique items auditioned	2.5	2.1	0.45	0.096
Total items auditioned	6.2	5.0	1.30	**0.036**
Duration (s)	53.1	30.9	22.21	**0.000**
Time per item (s)	13.0	9.4	3.59	**0.003**
Reauditions	3.7	2.9	0.85	**0.041**

The significant difference in *duration* observed between the two conditions leads to the conclusion that the novel glyph design, relative to the star glyph design, enabled subjects to complete trials more quickly. Likewise, the significant difference in *total*

items auditioned leads to the conclusion that the novel design also enabled subjects to complete trials while clicking on fewer icons. The larger *p*-value associated with *unique items auditioned*, however, makes it difficult to draw conclusions about a difference on this variable. The implications of these somewhat curious results are discussed further below.

5.6 Discussion

This section interprets the results presented above while focusing on performance differences between the glyph designs. It describes the ways in which the novel design is clearly superior to the star design, as well as ways in which the two perform equivalently. Following a comparison of the two designs on search task performance, some of the subjects' written responses from the post-test survey will be discussed. These responses provided insight into subjects' subjective impressions of the glyph designs. They also shed light on which visual features subjects found most useful in completing the task.

5.6.1 Differences in IR search task performance: three perspectives

The *t*-tests reported in Table 12 showed significant differences between the two glyph designs on four of the five variables examined. The only variable for which no difference was found was *unique items auditioned*. Given these results, what can be concluded about performance differences between the two glyph types? This question is considered from three angles:

1. the ability of the designs to convey the sound of not-yet-heard IRs
2. the ability of the designs to remind users of the sound of previously heard IRs
3. the impact of the designs on search speed

The novel design will be shown to clearly outperform the star design on the latter two points, despite a lack of evidence for superiority on the first.

5.6.1.1 Conveying the sound of not-yet-heard IRs

The *unique items auditioned* variable, in the context of the experiment, was intended to measure the ability of a glyph design to convey the acoustic character of a not-yet-heard IR. In an extreme case, if a glyph design were "perfect" in its communicative ability, subjects sufficiently well-trained in reverberation discrimination would be expected to complete the search task without clicking on any icons at all. Subjects would merely listen to the target, examine the nine glyphs, and would know immediately, from visual inspection, which one corresponded to the target. In such trials, the *unique items auditioned* variable would be zero. Given the complex nature of perceptual modeling of room acoustics, as well as the complex nature of glyph design, perfect performance was not anticipated. Nonetheless, the novel glyph design was expected to convey acoustic characteristics better than the star glyph design, and hence was expected to have a lower score on this variable.

While the mean value of *unique items auditioned* was indeed lower in the novel glyph design condition, the difference was not significant. In considering this result, it should be stressed that a non-significant result does not necessarily mean that no true difference exists between the groups means. Rather, a strict interpretation only asserts that no evidence for a difference could be found using this particular test. Had a test with higher statistical power been used, perhaps gained through a larger sample size, a difference may indeed have been detected.

Another obvious factor influencing the test results is the design of the experiment. The current experiment involved searches among sets of only nine IRs. This small number was chosen to keep trials short and make participation more appealing to subjects. This choice sacrificed some ecological validity, however, since the central goal of the research was to understand search interfaces for very large IR libraries, with hundreds or thousands of items. Had a more ecologically valid experimental design been chosen, where subjects were asked to locate a target within a much larger set of stimuli, statistical differences between the glyph types may have been more likely to emerge.

Clearly, more research is needed to better understand the ability of the novel design to visually convey the character of not-yet-heard IRs. Conclusions about other aspects the novel design's performance are easier to draw, however, and these are addressed below.

5.6.1.2 Serving as memory aids for previously heard IRs

Much as the *unique items auditioned* variable, discussed in the last section, can be considered a measure of a design's ability to initially convey an IR's sound, the *reauditions* variable is a measure of a design's ability to *remind* subjects of the sound of an IR. Each *reaudition* was a click on an IR that had already been heard, and each could have been avoided if the glyph design were better able to remind subjects of the IR's character. The t-test on *reauditions* showed that this variable was lower with the novel glyph design. This suggests that the novel design may be better able to function as an acoustic memory aid for previously heard IRs.

5.6.1.3 Search speed and trial duration

Star glyphs were associated with more *reauditions* than novel glyphs, and also with longer trial *durations.* Given the correlation between these two variables, one might posit a causal link between the two. Did star glyph trials take longer solely because more IR icons were clicked? Or were other factors at play in the differing trial *durations*? The variable *seconds per item* was designed to answer this question. If the difference in trial *duration* were solely due to additional IR listens, *seconds per item* would be expected to be similar in both conditions. Instead, however, *seconds per item* is lower with the novel glyphs. This indicates that the difference in trial *duration* was not simply due to a larger number of *reauditions.*

How was this additional time in the star glyph trials spent? Two explanations seem likely. The first is that subjects spent additional time in these trials consulting the legend. This would be expected, due to the low naturalness of the star glyph auditory-to-visual mappings. Since the mappings weren't natural, they needed to be held in short term memory, and, when they were forgotten, they would need to be refreshed through a glance at the legend. This explanation is supported by subjects' written comments. When asked to justify why they preferred the novel glyphs over the star glyphs, a majority of subjects cited having to consult the legend less often as the reason.

A second possible explanation for the larger *seconds per item* in the star glyph trials is that the additional time was spent on visual searches. As noted in section 5.2.3.3, vision science predicts that glyph searches will be faster when data attributes are mapped onto maximally distinct visual channels; that is, when visual channel separation is observed. Compared with the novel design, the star glyph design had only modest separation between visual channels. Theory would predict, then, that visual searches in star glyph displays would be relatively show, and that these slower searches would contribute to longer trial durations relative to the novel design. Slower visual searches, then, may also have contributed to longer trial *durations* with the star glyph design.

5.6.1.4 Summary

To summarize, then, the question of whether the novel glyph design "outperformed" the star glyph design can be considered from three perspectives. First, as revealed by the discussion of *unique items auditioned*, the present study cannot conclude that the two designs differ in their ability to convey the sound of a not-yet-heard IR. However, as discussed in the following section, evidence does suggest that they differ in their ability to remind users of the sounds of IRs heard previously. Finally, evidence also suggests that the two designs differ either in the amount of time required to interpret them, or in the speed with which they can be searched, or both. Thus, at the very least, the novel design appears to make IR searches more efficient in terms of reauditions and speed.

5.6.2 Subjective preferences for glyph designs

In a post-experiment survey, subjects were asked to identify their preferred glyph design and to justify their choice. Eleven out of twelve subjects said that they preferred the novel design. To support their answers, most cited the intuitiveness of the mappings, and the lack of need to consult a legend.

The sole subject to not prefer the novel glyphs included some useful feedback. Specifically, they noted that the visual feature of icon *distance* sometimes interfered with the legibility of other features. This is because icon size was used to cue distance: when icons were meant to appear distant, they were drawn smaller. The overall smaller size made other features such as *triangle length* and *oval width* difficult to resolve. Also, comparisons of *triangle length* and *oval width* were difficult when one icon was much smaller than the other. In these cases, some visual feature comparisons were simpler with the star glyph design. This comment highlights a legitimate shortcoming of the novel design which should be addressed in future work.

5.6.3 Which visual features were most helpful in carrying out the task?

An additional important question about the novel design concerns the relative utility of each of the four visual features for completing the search task. If subjects relied more heavily on certain features than others, this may suggest areas of weakness in the design. Visual features which were less useful should perhaps be reconsidered or modified. As above, some insight into relative utility of different features is given in subjects' written responses on the post-experiment survey.

In their comments, several subjects explained that their first action, upon hearing the target IR, was to narrow down their search to a small number of candidate icons by filtering on one single visual feature. The feature most often cited for this initial filtering was *colour* (mapped to auditory brightness). That is, subjects heard the target, evaluated its auditory brightness, mapped this brightness to a *colour*, and then focused their attention only on those icons matching this *colour*.

This use of colour filtering as a first step in visual search is consistent with findings in vision science. When faced with similar tasks, subjects are known to make use of so-called “guiding features” to narrow visual searches. Multiple features including icon shape, size, and depth are all capable of functioning as “guides” to varying degrees, but colour is thought to be one of the most effective (Wolfe & Horowitz, 2004).

Given that many subjects initiated the IR search task by filtering on one specific visual feature, future refinements of the novel design might ensure that all visual features employed were equally capable of acting in this “guiding” role. If other visual features were equally effective guides, search performance within certain sets of IRs, particularly within sets of equal auditory brightness, might be improved.

Of course, while this discussion suggests an explanation for subjects' reliance on icon colour rooted in visual perception, it is possible that the explanation involves auditory perception as well. That is, perhaps attending to variations in the timbral brightness of stimuli was cognitively easier for subjects than attending to variations in the other three attributes. In this way, the attributes' relative salience, in the minds of sound engineers, may play a role in determining search strategies. Future work on IR glyph design would thus also benefit from a better understanding of attribute salience, and of the ease with which comparisons on each attribute can be made.

5.7 Conclusions

Having elaborated a perceptual model of natural reverberation in earlier chapters, this chapter attempted to develop data visualization techniques to facilitate the task of searching within large IR libraries. Following a review of relevant vision science principles, a visualization strategy for natural IRs was proposed called the novel glyph design. In an empirical study, the novel design was shown to enable users to complete IR searches more quickly and efficiently relative to a baseline visualization. Several visual shortcomings of the novel design were also highlighted. These shortcomings should be addressed in future work.

6 CONCLUSIONS

6.1 Research questions

This dissertation aimed to answer a series of questions related to the design of interfaces for perceptual browsing of large convolution reverb impulse responses libraries.

6.1.1 Perceptual structure of impulse response libraries

The first question concerned the perceptual structure of impulse response libraries. What are the primary dimensions of perceptual variation in such libraries, and what are the acoustic correlates of these dimensions? More succinctly, what is a useful perceptual model for libraries of convolution reverb IRs?

This question was answered through a series of investigations reported in chapters two through five. Chapter two established that the perceptual model in ISO-3382-1, a four-dimensional model including the attributes of reverberance, clarity, ASW and LEV, appeared to characterize fairly well the perceived differences between concert venues heard by audience members. Chapter three investigated whether this model was also adequate for describing the variations heard by trained sound engineers in a library of loudness-matched acoustic impulse responses. Specifically, it explored whether any additional attributes could be identified for this population of listeners in this particular listening context. An analysis of the words used by engineers to describe differences between IRs suggested two potential additional attributes: brightness and source distance.

Chapter 4 studied these two attributes in detail and concluded that the source distance attribute was correlated with the same objective signal features as ISO-3382-1 clarity, suggesting that source distance was not a distinct attribute but rather a different name for an existing attribute. However, as the sound engineers in the study did appear to favour distance-related words to clarity-related words when describing variations in this attribute, this finding was notable and may have potential applications in the design of maximally intuitive search parameter names.

Chapter 5 initially presented a six-attribute model to characterize the perceptual qualities of IRs but noted that the signal features predicting reverberance and LEV, and clarity and source distance, respectively, were highly correlated. These correlations suggested that the IRs could be described by a more compact four-dimensional model with little loss of information. This research, then, suggests that, in the library under investigation, a four-dimensional model was adequate for characterizing the most salient perceptual dimensions of IRs. The most natural names for these four dimensions appeared to be reverberance, ASW, brightness and source distance.

With respect to acoustic correlates, the literature review found the *EDT* feature to be a good predictor of reverberance, and *sASW,* a feature derived from a binaural auditory model, to be a good predictor of ASW. Ratings of source distance were best predicted by a sigmoid-transformed version of the ISO C_{80}. Brightness ratings were best predicted by a spectral slope computed from a time-windowed IR plotted on perceptually scaled axes.

6.1.2 Visual representations of IRs

The second question concerned visual representations of impulse responses within a library. Can appropriate visualizations of IRs assist users in locating items with particular perceptual characteristics?

This question was addressed in Chapter 5. Here, an experiment showed that, indeed, two different IR visualization strategies were associated with performance differences on an IR search task. Specifically, a novel application-specific visualization reduced search times by a factor of two, relative to a general-purpose “star glyph” visualization.

6.2 Contributions and future work

Having presented answers to the two central research questions of the dissertation, this section will discuss how these primary findings, along with other incidental contributions, relate to broader literature on reverberation perception and reverberation effect user interface design.

6.2.1 IR spectral slope as a measure of reverberation timbre

Section 4.4.3.2 investigated correlations between judgements of reverberation brightness and a collection of objective signal features. Features drawn from the existing literature on reverberation timbre were examined (e.g. EBL, TR_{late}) as well as a novel feature related to the IR spectral slope. The novel spectral slope feature outperformed all other IR-based measures in its predictive ability. This is a notable finding. Timbral attributes such as brightness have emerged in many exploratory studies as important sources of perceptual variation in acoustic reverberation, yet their objective predictors remain under-investigated. Although features similar to IR spectral slope have been proposed (Chourdakis & Reiss, 2019), no work besides this dissertation to the knowledge of the author has compared slope-related features to earlier timbral features in terms of their perceptual relevance.

Better IR-based predictors of reverberation timbre are needed by the room acoustics community (J. S. Bradley, 2011). The results presented in Chapter four suggest that the IR spectral slope may be an improvement over state-of-the-art timbral features. Future studies should attempt to confirm this finding.

6.2.2 Visualizations of perceptual attributes (IR glyph design)

Visual representations of reverberation aimed at facilitating searches within libraries of IRs are a relatively unexplored topic. This dissertation contributed one such visualization. The experiment in chapter five showed that this visual design was relatively easy to interpret and that it outperformed a baseline visualization on an IR

search task. This dissertation also contributed a methodology for evaluating novel IR visualizations, reported in section 5.4.4.2.

The novel visualization could certainly be improved. In particular, the proposed visual mappings for the attributes of ASW and source distance were frequently misunderstood. Future research should attempt to develop mappings for these attributes that are less ambiguous. Additionally, the design should be compared in its search efficacy with other IR-specific visualization techniques found in the literature (Monks et al., 2000; Stettner & Greenberg, 1989). Both earlier designs used perceptual models that were simpler than the one proposed herein, and both were created for a different application domain (acoustical design, rather than audio production). Nonetheless, it remains possible that these visualizations might either be directly useful in IR search interfaces or might be useful for inspiring improvements to the novel design.

6.2.3 Intuitive language for perceptual attributes of natural reverberation

One small contribution, introduced in section 3.3, concerns intuitive language for describing attributes of natural reverberation within sound engineering communities. When exploring relationships between freely elicited descriptions of reverberant stimuli and objective measurements on these stimuli, subjects seemed to associate differences in *EDT* with words related to time ("shorter"/"longer" decay time) and differences in C_{80} with words related to distance ("closer"/"farther" sound source). These word choices diverge subtly from the descriptive language used in the acoustics research community, which associates *EDT* with changes in "reverberance" and C_{80} with changes in "clarity". This preference for temporal- and distance-related language with experimental subjects, however, suggests that, for sound engineers, the former sorts of words might be more intuitive descriptions for variations in *EDT* and C_{80}.

6.2.4 Analysis of algorithmic reverb preset names

In a similar vein, the descriptive labels associated with algorithmic reverb presets, presented in section 3.1.1.2, also present a modest contribution related to descriptive language for reverberation. In contrast to the previous section, the labels presented here refer to synthetic rather than natural reverb and were applied by reverb designers rather than users of reverberation effects. This data complements work by Seetharam and Pardo on perceptual descriptions of algorithmic presets by non-experts (2014) and Garland and Ronan's analysis of algorithmic reverb control parameter names (2021). By summarizing the descriptive language used by reverb designers in a library spanning three decades, this data provide some insight into preset naming conventions within the community.

6.2.5 Objective correlates of algorithmic preset labels

In addition to identifying common descriptive terms for algorithmic presets, section 3.1.3 also provides some objective correlates for these terms. In particular, the weights shown in Figure 10 give an indication of the features most strongly associated with each label. These relationships are not always intuitive. For example, the *dense* model has a positive weight on *IACC* and a negative weight on *pASW*. This suggests that the "dense" label was associated with high correlation between the left and right channels and narrow source images. Conversely, the *rich* model, with negative weights on both *IACC* and *IACC* late, was associated with low inter-channel correlation. The *vocal* model seemed to be defined by bright reverb timbre (e.g. low *early bass level* and high *late treble ratio*) as well as a low C_{80}. Other results are unsurprising, such as the negative weight on NED mixing time in the *chamber* model. This suggests that presets emulating the sounds of reverb chambers were partially defined by short mixing times and an early diffuse field onset. These results provide a starting point for future machine learning efforts aimed at automatically applying labels to reverb IRs or recommending IRs for specific source material.

6.2.6 Ecological validation of dynamic query-based IR browsing systems

Perhaps the most obvious contribution of this work is the idea of searching IR libraries via dynamic queries on their perceptual attributes. A dynamic query-like interface, of the type presented in section 1.3 and using the perceptual model and glyph design derived in Chapter 5, would seem to offer advantages over the more standard search interface design presented in the introduction. Such an interface would not only allow searches to be carried out in an intuitive perceptual space, but would also allow searches in two simultaneous modalities: filter sliders would control which IRs were visible in the display, while human vision would control which of the displayed IRs were most visually salient, through selective attention to separable visual features such as icon colour, shape and size.

The final and most important area for future study related to this work involves the design of a fully functional dynamic query-based perceptual search interface, and the evaluation of this interface in ecologically valid contexts. Such work might involve a usability comparison of the novel interface with more traditional interfaces on controlled search tasks (c.f., Benson & Woszczyk, 2012). Or it might involve more subjective longitudinal evaluations carried out by inviting mixing engineers to use the interface in their work and provide periodic feedback.

Indeed, one search interface drawing on this work's results has been developed in parallel with this research. A screenshot of this interface, featuring the perceptual model and glyph design developed in Chapter 5, is visible in Figure 39. This interface is being used as a browser for the Spacebuilder library within an experimental convolution reverb plugin. Evaluations of the interface's usability in audio production workflows will be forthcoming.

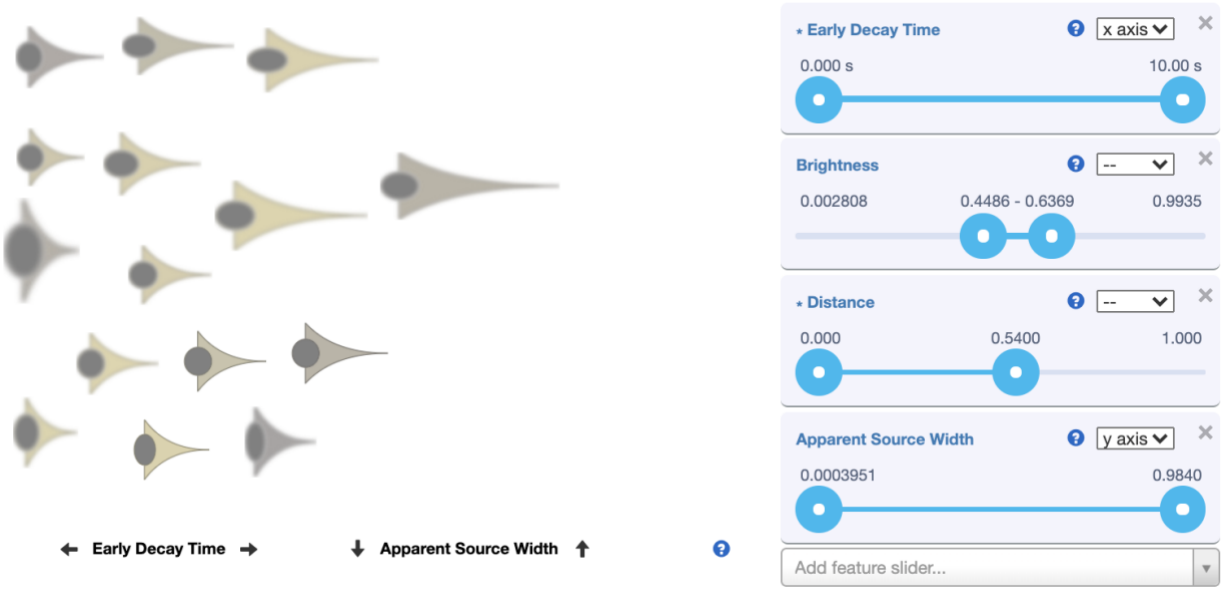


Figure 39 Prototype perceptual search interface

In 1991, Anders Gade closed a paper at the 90th convention of the Audio Engineering Society with the following thought, musing about the eventual possibility of convolution reverb as common studio effect:

> Finally, the still increasing computer capabilities makes one wonder, how soon the day will come when sound effects processors will be based on measured impulse responses from real rooms instead of on simple reverberation algorithms [...] ! (p. 12)

Eight years later, the first commercial convolution reverberator was released (Sony, 1999). The fact that convolution effects were fanciful as recently as 1991, and only widely available in the early 2000's, underscores the relative novelty of these effects, and hence the novelty of the central research problems addressed in this work. Given that search interfaces for IR libraries have only existed for two decades, it is perhaps unsurprising that few design idioms for such interfaces have been proposed!

Still, potential applications for high-quality multichannel reverberation are multiplying, from "virtual acoustic" settings in which natural and synthetic reverberation are combined for musical ends (Ko & Woszczyk, 2018), to telepresence and virtual reality, where believable synthesized acoustics must align with imagined or created virtual spaces. If, in these contexts, reverberation is to be generated through convolution, each represents a use case for a scalable, context sensitive, perceptually guided IR search interface. As IR libraries continue to grow, and as they find novel uses both inside and outside of the recording studio, intuitive perceptual search interfaces, well-adapted to large collections, will become increasingly important.

7 WORKS CITED

Abel, J. S., Bryan, N. J., Huang, P. P., Kolar, M. A., & Pentcheva, B. V. (2010). *Estimating Room Impulse Responses from Recorded Balloon Pops* [Paper presentation]. 129th Convention of the Audio Engineering Society, San Francisco, CA.

Abel, J. S., & Huang, P. (2006). *A Simple, Robust Measure of Reverberation Echo Density* [Paper presentation]. 121st Convention of the Audio Engineering Society, San Francisco, CA.

Abel, J. S., & Huang, P. (2007). *Aspects of Reverberation Echo Density* [Paper presentation]. 123rd Convention of the Audio Engineering Society, New York, NY.

Ahlberg, C., & Shneiderman, B. (1994). Visual information seeking using the FilmFinder. *Conference Companion on Human Factors in Computing Systems - CHI '94*, 433–434. https://doi.org/10.1145/259963.260431

Ahlberg, C., & Shneiderman, B. (2003). Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays. In B. B. Bederson & B. Shneiderman (Eds.), *The Craft of Information Visualization* (pp. 7–13). Morgan Kaufmann. https://doi.org/10.1016/B978-155860915-0/50004-4

Ahlberg, C., Williamson, C., & Shneiderman, B. (1992). Dynamic Queries for Information Exploration: An Implementation and Evaluation. In P. Bauerfeld, J. Bennett, & G. Lynch (Eds.), *Proceedings of the ACM CHI 92 Human Factors in Computing Systems Conference* (pp. 619–626).

https://www.acm.org/pubs/articles/proceedings/chi/142750/p619-ahlberg/p619-ahlberg.pdf

Anderson, E. (1957). A Semigraphical Method for the Analysis of Complex Problems. *Proceedings of the National Academy of Sciences of the United States of America*, *43*(10), 923–927.

Atal, B. S., Schroeder, M. R., & Sessler, G. M. (1965). Subjective reverberation time and its relation to sound decay. *Proceedings of Th Fifth International Congress on Acoustics*, *1b*, Paper G32.

Audio Ease. (2012). *Altiverb* (Version 7) [Computer software]. https://www.audioease.com/altiverb/

Bair, E., Hastie, T., Paul, D., & Tibshirani, R. (2006). Prediction by Supervised Principal Components. *Journal of the American Statistical Association*, *101*(473), 119–137. https://doi.org/10.1198/016214505000000628

Barron, M. (1971). The subjective effects of first reflections in concert halls—The need for lateral reflections. *Journal of Sound and Vibration*, *15*(4), 475–494. https://doi.org/10.1016/0022-460X(71)90406-8

Barron, M. (1988). Subjective Study of British Symphony Concert Halls. *Acta Acustica United with Acustica*, *66*(1), 1–14.

Barron, M. (2005). Using the standard on objective measures for concert auditoria, ISO 3382, to give reliable results. *Acoustical Science and Technology*, *26*(2), 162–169. https://doi.org/10.1250/ast.26.162

Barron, M., & Marshall, A. (1981). Spatial impression due to early lateral reflections in concert halls: The derivation of a physical measure. *Journal of Sound and Vibration*, *77*(2), 211–232.

Becker, J. (2002). Spectral and temporal contribution of different signals to ASW analysed with binaural hearing models. *Proceedings of the Forum Acusticum 2002*, 1–6.

Ben-Hador, R., & Neoran, I. (2004). *Capturing manipulation and reproduction of sampled acoustic impulse responses* [Paper presentation]. 117th Convention of the Audio Engineering Society, San Francisco, CA. http://www.aes.org/e-lib/browse.cfm?elib=12889

Benson, D., & Woszczyk, W. (2012). *Searching Impulse Response libraries using Room Acoustic Descriptors* [Poster presentation]. 133rd Convention of the Audio Engineering Society, San Francisco, CA. https://www.aes.org/e-lib/online/browse.cfm?elib=16500

Beranek, L. L. (1962). *Music, Acoustics & Architecture*. John Wiley & Sons, Inc.

Berg, J., & Rumsey, F. (2006). Identification of Quality Attributes of Spatial Audio by Repertory Grid Technique. *J. Audio Eng. Soc.*, *54*(5), 15.

Blauert, J., & Lindemann, W. (1986). Auditory spaciousness: Some further psychoacoustic analyses. *The Journal of the Acoustical Society of America*, *80*(2), 533–542. https://doi.org/10.1121/1.394048

Blesser, B. (2001). An Interdisciplinary Integration of Reverberation. *NEW YORK*, 18.

Bliefnick, J. M. (2016). *Investigation of Subjective Perception & Objective Metrics of Acoustic Room Diffusion*. University of Nebraska-Lincoln.

Borgo, R., Kehrer, J., Chung, D. H. S., Maguire, E., Laramee, R. S., Hauser, H., Ward, M., & Chen, M. (2012). Glyph-based Visualization: Foundations, Design Guidelines, Techniques and Applications. *Eurographics 2013 - State of the Art Reports*, 25 pages. https://doi.org/10.2312/CONF/EG2013/STARS/039-063

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(7), 1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2

Bradley, J. S. (2011). Review of objective room acoustics measures and future needs. *Applied Acoustics*, *72*(10), 713–720. https://doi.org/10.1016/j.apacoust.2011.04.004

Bradley, J. S., & Soulodre, G. A. (1995). Objective measures of listener envelopment. *The Journal of the Acoustical Society of America*, *98*(5), 2590–2597. https://doi.org/10.1121/1.413225

Breebaart, J., van de Par, S., & Kohlrausch, A. (2001). Binaural processing model based on contralateral inhibition. I. Model structure. *The Journal of the Acoustical Society of America*, *110*(2), 1074. https://doi.org/10.1121/1.1383297

Brinkmann, F., Aspöck, L., Ackermann, D., Lepa, S., Vorländer, M., & Weinzierl, S. (2019). A round robin on room acoustical simulation and auralization. *The Journal of the Acoustical Society of America*, *145*(4), 2746–2760. https://doi.org/10.1121/1.5096178

Bronkhorst, A. W., & Houtgast, T. (1999). Auditory distance perception in rooms. *Nature*, *397*(6719), 517–520. https://doi.org/10.1038/17374

Bryan, N. J., & Abel, J. S. (2010, November). *Methods for Extending Room Impulse Responses Beyond Their Noise Floor* [Paper presentation]. 129th Convention of the Audio Engineering Society, San Francisco, CA. https://www.aes.org/e-lib/browse.cfm?elib=15590

Cannam, C., Landone, C., & Sandler, M. (2010). Sonic visualiser: An open source application for viewing, analysing, and annotating music audio files. *Proceedings of the International Conference on Multimedia - MM '10*, 1467–1468. https://doi.org/10.1145/1873951.1874248

Canty, A., & Ripley, B. (2021). *boot: Boostrap R (S-Plus) Functions* (R package version 1.3-28) [Computer software].

Case, A. (2012). *Sound FX: Unlocking the Creative Potential of Recording Studio Effects* (0 ed.). Routledge. https://doi.org/10.4324/9780080548968

Chambers, J. M., Cleveland, W. S., Kleiner, B., & Tukey, P. A. (1983). *Graphical methods for data analysis*. CRC Press/Taylor & Francis Group.

Chernyak, R. I., & Dubrovski, N. A. (1968). Pattern of the noise images and the binaural summation of loudness for the different interaural correlation of noise. *Proc. 6th Int. Congr. Acoust., A-3-12*, A53–A56.

Chourdakis, E. T., & Reiss, J. D. (2019). *Tagging and retrieval of room impulse responses using semantic word vectors and perceptual measures of reverberation* [Paper presentation]. 146th Convention of the Audio Engineering Society, Dublin, IE.

Cox, T. J., & D'Antonio, P. (2017). *Acoustic absorbers and diffusers: Theory, design and application* (Third edition). CRC Press/Taylor & Francis Group.

Cremer, L. 1905-1990., & Müller, H. A. (1982). *Principles and applications of room acoustics*. Applied Science; WorldCat.org.

Dau, T., Püschel, D., & Kohlrausch, A. (1996). A quantitative model of the "effective" signal processing in the auditory system. I. Model structure. *The Journal of the Acoustical Society of America*, *99*(6), 3615–3622.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap Methods and Their Applications*. Cambridge University Press.

Durlach, N. I. (1960). Note on the equalization and cancellation theory of binaural masking level differences. *The Journal of the Acoustical Society of America*, *32*(8), 1075–1076.

Farina, A. (2000). *Simultaneous Measurement of Impulse Response and Distortion with a Swept-Sine Technique* [Paper presentation]. 108th Convention of the Audio Engineering Society, Paris, FR.

Field, A., Miles, J., & Field, Z. (2012). *Discovering Statistics Using R*. Sage Publications Ltd.

Flux:: Software Engineering. (2021). *Spat Revolution* (21.4.0.50030) [Computer software]. Flux:: Software Engineering. https://www.flux.audio/

Friedman, J., Hastie, T., & Tibshirani, R. (2009). *The Elements of Statistical Learning* (2nd ed.). Springer. http://statweb.stanford.edu/~tibs/book/preface.ps

Froehlich, B., Hochstrate, J., Skuk, V., & Huckauf, A. (2006). The GlobeFish and the GlobeMouse: Two new six degree of freedom input devices for graphics applications. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 191–199. https://doi.org/10.1145/1124772.1124802

Fuchs, J., Isenberg, P., Bezerianos, A., Fischer, F., & Bertini, E. (2014). The Influence of Contour on Similarity Perception of Star Glyphs. *IEEE Transactions on Visualization and Computer Graphics*, *20*(12), 2251–2260. https://doi.org/10.1109/TVCG.2014.2346426

Fuchs, J., Isenberg, P., Bezerianos, A., & Keim, D. (2017). A Systematic Review of Experimental Studies on Data Glyphs. *IEEE Transactions on Visualization and Computer Graphics*, *23*(7), 1863–1879. https://doi.org/10.1109/TVCG.2016.2549018

Gade, A. C. (1991). *Assessment of sound quality in auditoria* [Paper presentation]. 90th Convention of the Audio Engineering Society, Paris.

Gardner, W. G. (2002). Reverberation Algorithms. In M. Kahrs & K. Brandenburg (Eds.), *Applications of Digital Signal Processing to Audio and Acoustics* (Vol. 437, pp. 85–131). Kluwer Academic Publishers. https://doi.org/10.1007/0-306-47042-X_3

Garland, K., & Ronan, M. (2021). *Defining reverberation plugin structure: A comparative exploration of system design and expert knowledge in an audio education context* [Paper presentation]. 151st Convention of the Audio Engineering Society, Online.

Golland, P., & Fischl, B. (2003). Permutation Tests for Classification: Towards Statistical Significance in Image-Based Studies. In C. Taylor & J. A. Noble (Eds.), *Proc. IPMI'03: The 18th International Conference on Information Processing in Medical Imaging* (pp. 330–341). Springer.

Grey, J. M. (1977). Multidimensional perceptual scaling of musical timbres. *The Journal of the Acoustical Society of America, 61*(5), 1270–1277. https://doi.org/10.1121/1.381428

Griesinger, D. (2015). Acoustic quality, proximity, and localization in concert halls: The role of harmonic phase alignment. *Psychomusicology: Music, Mind, and Brain, 25*(3), 339–344. https://doi.org/10.1037/pmu0000116

Hawkes, R. J., & Douglas, H. (1971). Subjective Acoustic Experience in Concert Auditoria. *Acta Acustica United with Acustica, 24*(5), 235–250.

Hidaka, T., Beranek, L. L., & Okano, T. (1995). Interaural cross-correlation, lateral fraction, and low- and high-frequency sound levels as measures of acoustical quality in concert halls. *The Journal of the Acoustical Society of America, 98*(2), 988–1007. https://doi.org/10.1121/1.414451

Inglis, S. (2020). *Reverb: What Do All Those Knobs Do?* Sound on Sound. https://www.soundonsound.com/techniques/reverb-what-do-all-those-knobs-do

International Organization for Standardization. (2008). *Acoustics—Measurement of room acoustic parameters—Part 2: Reverberation time in ordinary rooms (ISO 3382-2:2008)*. https://www.iso.org

International Organization for Standardization. (2009). *Acoustics—Measurement of room acoustic parameters—Part 1: Performance spaces (ISO 3382-1:2009)*. https://www.iso.org

International Telecommunications Union. (2011). *Recommendation ITU-R BS.1770-2—Algorithms to measure audio programme loudness and true-peak audio level*. https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.1770-2-201103-S!!PDF-E.pdf

Jeon, J. Y., Jang, H. S., Kim, Y. H., & Vorländer, M. (2013). Subjective and objective evaluations of scattered sounds in concert halls. *Proceedings of the International Symposium on Room Acoustics (ISRA 2013)*, Paper no. 077.

Jordan, V. L. (1970). Acoustical Criteria for Auditoriums and Their Relation to Model Techniques. *The Journal of the Acoustical Society of America*, *47*(2A), 408–412. https://doi.org/10.1121/1.1911535

Jot, J.-M., Cerveau, L., & Warusfel, O. (1997). *Analysis and synthesis of room reverberation based on a statistical time-frequency model* [Paper presentation]. 103rd Convention of the Audio Engineering Society, New York, NY.

Kahle, E. (1995). *Validation d'un modèle objectif de la perception de la qualité acoustique dans un ensemble de salles de concerts et d'opéras* [Doctoral dissertation, Université du Maine]. https://kahle.be/articles/These_EK.pdf

Karve, A., & Gleicher, M. (2007). Glyph-based Overviews of Large Datasets in Structural Bioinformatics. *11th International Conference Information Visualization - Supplements (IV '07)*, 1–6. https://doi.org/10.1109/IV.2007.150

Keet, W. V. (1968). The influence of early lateral reflections on the spatial impression. *Proc. 6th Int. Cong. Acoust., Tokyo*, *E-2-4*, E53–E56.

Ko, D., & Woszczyk, W. (2018). Virtual Acoustics for Musicians: Subjective Evaluation of a Virtual Acoustic System in Performance of String Quartets. *J. Audio Eng. Soc.*, *66*(9), 712–723. https://doi.org/10.17743/jaes.2018.0038

Lakatos, S. (2000). A common perceptual space for harmonic and percussive timbres. *Perception & Psychophysics*, *62*(7), 1426–1439. https://doi.org/10.3758/BF03212144

Larsen, E., Iyer, N., Lansing, C. R., & Feng, A. S. (2008). On the minimum audible difference in direct-to-reverberant energy ratio. *The Journal of the Acoustical Society of America*, *124*(1), 450–461. https://doi.org/10.1121/1.2936368

Lee, D., & Cabrera, D. (2010). Effect of listening level and background noise on the subjective decay rate of room impulse responses: Using time-varying loudness to model reverberance. *Applied Acoustics*, *71*(9), 801–811. https://doi.org/10.1016/j.apacoust.2010.04.005

Lee, D., van Dorp Schuitman, J., Cabrera, D., Qiu, X., & Burnett, I. (2017). Comparison of psychoacoustic-based reverberance parameters. *The Journal of the Acoustical Society of America*, *142*(4), 1832–1840. https://doi.org/10.1121/1.5005508

Legendre, P., & Legendre, L. (2012). *Numerical Ecology*. Elsevier.

Lexicon. (2021). *PCM Native Reverb Plugin* (1.3.10) [Computer software]. Lexicon. https://lexiconpro.com/en/products/pcm-native-reverb-plug-in-bundle

Lindau, A. (2012). Perceptual Evaluation of Model- and Signal-Based Predictors of the Mixing Time in Binaural Room Impulse Responses. *Journal of the Audio Engineering Society*, *60*(11), 887–898.

Lindemann, W. (1986). Extension of a binaural cross-correlation model by contralateral inhibition. I. Simulation of lateralization for stationary signals. *The Journal of the Acoustical Society of America*, *80*(6), 1608–1622. https://doi.org/10.1121/1.394325

Lokki, T., Pätynen, J., Kuusinen, A., & Tervo, S. (2012). Disentangling preference ratings of concert hall acoustics using subjective sensory profiles. *The Journal of the Acoustical Society of America*, *132*(5), 3148–3161.

Lokki, T., Pätynen, J., Kuusinen, A., Vertanen, H., & Tervo, S. (2011). Concert hall acoustics assessment with individually elicited attributes. *The Journal of the Acoustical Society of America*, *130*(2), 835–849.

Lu, Y.-C., & Cooke, M. (2010). Binaural Estimation of Sound Source Distance via the Direct-to-Reverberant Energy Ratio for Static and Moving Sources. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*(7), 1793–1805. https://doi.org/10.1109/TASL.2010.2050687

Marshall, A. H. (1967). A note on the importance of room cross-section in concert halls. *Journal of Sound and Vibration*, *5*(1), 100–112. https://doi.org/10.1016/0022-460X(67)90181-2

Marshall, A. H., & Barron, M. (2001). Spatial responsiveness in concert halls and the origins of spatial impression. *Applied Acoustics*, *62*(2), 91–108. https://doi.org/10.1016/S0003-682X(00)00050-5

Masiero, B. (2007). *AcMus—Room Acoustic Parameters* (1.0.0.0) [Computer software]. https://www.mathworks.com/matlabcentral/fileexchange/11392-acmus-room-acoustic-parameters

Mason, R., Brookes, T., & Rumsey, F. (2005). Frequency dependency of the relationship between perceived auditory source width and the interaural cross-correlation coefficient for time-invariant stimuli. *The Journal of the Acoustical Society of America*, *117*(3), 1337–1350. https://doi.org/10.1121/1.1853113

McAdams, S. (2019). The Perceptual Representation of Timbre. In K. Siedenburg, C. Saitis, S. McAdams, A. N. Popper, & R. R. Fay (Eds.), *Timbre: Acoustics, Perception, and Cognition* (pp. 23–57). Springer International Publishing. https://doi.org/10.1007/978-3-030-14832-4_2

McDougall, S. J. P., de Bruijn, O., & Curry, M. B. (2000). Exploring the Effects of Icon Characteristics on User Performance: The Role of Icon Concreteness, Complexity, and Distinctiveness. *Journal of Experimental Psychology: Applied, 6*(4), 291–306. https://doi.org/10.1037/1076-898X.6.4.291

Monks, M., Oh, B. M., & Dorsey, J. (2000). Audioptimization: Goal-based acoustic design. *IEEE Computer Graphics and Applications, 20*(3), 76–90. https://doi.org/10.1109/38.844375

Morimoto, M. (2002). The Relation Between Spatial Impression and the Precedence Effect. *Proceedings of the Eighth International Conference on Auditory Display (ICAD 2002)*. https://www.icad.org/websiteV2.0/Conferences/ICAD2002/proceedings/Morimoto1.pdf

Morimoto, M., Iida, K., & Furue, Y. (1993). Relation between auditory source width in various sound fields and degree of interaural cross-correlation. *Applied Acoustics, 38*(2–4), 291–301. https://doi.org/10.1016/0003-682X(93)90057-D

Morimoto, M., Jinya, M., & Nakagawa, K. (2007). Effects of frequency characteristics of reverberation time on listener envelopment. *The Journal of the Acoustical Society of America, 122*(3), 1611–1615. https://doi.org/10.1121/1.2756164

Morimoto, M., Sugiura, S., & Iida, K. (1994). Relation between auditory source width in various sound fields and degree of interaural cross-correlation: Confirmation by

constant method. *Applied Acoustics*, *42*(3), 233–238. https://doi.org/10.1016/0003-682X(94)90111-2

Neider, M. B., & Zelinsky, G. J. (2006). Searching for camouflaged targets: Effects of target-background similarity on visual search. *Vision Research*, *46*(14), 2217–2235. https://doi.org/10.1016/j.visres.2006.01.006

Nishimura, A., & Sasaki, M. (2004). Absolute cues for auditory distance in front and lateral directions. *Acoustical Science and Technology*, *25*(2), 127–135. https://doi.org/10.1250/ast.25.127

Nolan, M., Berzborn, M., & Fernandez-Grande, E. (2020). Isotropy in decaying reverberant sound fields. *The Journal of the Acoustical Society of America*, *148*(2), 1077–1088. https://doi.org/10.1121/10.0001769

Norman, D. A. (1988). *The Psychology of Everyday Things* (pp. xi, 257). Basic Books.

Nuñez, J. R., Anderton, C. R., & Renslow, R. S. (2018). Optimizing Colormaps With Consideration for Color Vision Deficiency to Enable Accurate Interpretation of Scientific Data. *PLOS ONE*, *13*(7). https://doi.org/10.1371/journal.pone.0199239

Okano, T., Beranek, L. L., & Hidaka, T. (1998). Relations among interaural cross-correlation coefficient (IACCE), lateral fraction (LFE), and apparent source width (ASW) in concert halls. *The Journal of the Acoustical Society of America*, *104*(1), 255–265. https://doi.org/10.1121/1.423955

Parkin, P. H., & Morgan, K. (1965). "Assisted Resonance" in the Royal Festival Hall, London. *Journal of Sound and Vibration*, *2*(1), 74–85.

Pätynen, J., Pulkki, V., & Lokki, T. (2008). Anechoic Recording System for Symphony Orchestra. *Acta Acustica United with Acustica*, *94*(6), 856–865. https://doi.org/10.3813/AAA.918104

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, *49*(12), 1373–1379.

Peeters, G., Giordano, B. L., Susini, P., Misdariis, N., & McAdams, S. (2011). The timbre toolbox: Extracting audio descriptors from musical signals. *The Journal of the Acoustical Society of America*, *130*(5), 2902–2916.

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Ramakrishnan, R., & Grewal, A. (2008). Reverberation rooms and spatial uniformity. *Canadian Acoustics*, *36*(3), 28–29.

Reichardt, W., Alim, O. A., & Schmidt, W. (1975). Definition and Basis of Making an Objective Evaluation to Distinguish Between Useful and Useless Clarity Defining Musical Performances. *Acta Acustica United with Acustica*, *32*(3), 126–137.

Robinson, P. W., Pätynen, J., Lokki, T., Suk Jang, H., Yong Jeon, J., & Xiang, N. (2013). The role of diffusive architectural surfaces on auditory spatial discrimination in performance venues. *The Journal of the Acoustical Society of America*, *133*(6), 3940–3950. https://doi.org/10.1121/1.4803846

Robinson, P. W., Walther, A., Faller, C., & Braasch, J. (2013). Echo thresholds for reflections from acoustically diffusive architectural surfaces. *The Journal of the Acoustical Society of America*, *134*(4), 2755–2764. https://doi.org/10.1121/1.4820890

Rogers, B., & Graham, M. (1979). Motion Parallax as an Independent Cue for Depth Perception. *Perception*, *8*(2), 125–134. https://doi.org/10.1068/p080125

Ruckdeschel, P., & Kohl, M. (2014). General Purpose Convolution Algorithm in S4 Classes by Means of FFT. *Journal of Statistical Software*, *59*(4), 1–25.

Ruckdeschel, P., Kohl, M., Stabla, T., & Camphausen, F. (2006). S4 Classes for Distributions. *R News*, *6*(2), 2–6.

Rumsey, F. (2002). Spatial Quality Evaluation for Reproduced Sound: Terminology, Meaning, and a Scene-Based Paradigm. *J. Audio Eng. Soc.*, *50*(9), 651–666.

Rumsey, F., & Berg, J. (2001). Verification and correlation of attributes used for describing the spatial quality of reproduced sound. *Proc. of the AES 19th Int. Conf.*, 233–251.

Rumsey, F., Zielinski, S., Jackson, P., Dewhirst, M., Conetta, R., George, S., Bech, S., & Meares, D. (2008). *QESTRAL (Part 1): Quality Evaluation of Spatial Transmission and Reproduction using an Artificial Listener* [Paper presentation]. 125th Convention of the Audio Engineering Society, San Francisco, CA.

Sabine, W. C. (1900). *Reprints from the American Architect on Architectural Acoustics Part 1. - Reverberation*. American Architect. https://www.google.ca/books/edition/Reprints_from_the_American_Architect_on/XzbYfkU7Qf4C

Savioja, L., & Svensson, U. P. (2015). Overview of geometrical room acoustic modeling techniques. *The Journal of the Acoustical Society of America*, *138*(2), 708–730. https://doi.org/10.1121/1.4926438

Schroeder, M., Atal, B., & Sessler, G. (1966). Acoustical measurements in Philharmonic Hall (New York). *Journal of the Acoustical Society of America*, *40*(2), 434–440.

Schroeder, M. R. (1965). New Method of Measuring Reverberation Time. *Journal of the Acoustical Society of America*, *37*, 409–412. https://doi.org/10.1121/1.1909343

Schroeder, M. R., Gottlob, D., & Siebrasse, K. F. (1974). Comparative study of European concert halls: Correlation of subjective preference with geometric and acoustic

parameters. *The Journal of the Acoustical Society of America*, *56*(4), 1195–1201. https://doi.org/10.1121/1.1903408

Schubert, E., Wolfe, J., & Tarnopolsky, A. (2004). Spectral Centroid and Timbre in Complex, Multiple Instrumental Textures. *Proceedings of the 8th International Conference on Music Perception & Cognition*, 654–657.

Seetharaman, P., & Pardo, B. (2014). Crowdsourcing a Reverberation Descriptor Map. *Proceedings of the 22nd ACM International Conference on Multimedia*, 587–596. https://doi.org/10.1145/2647868.2654908

Shigemizu, D., Akiyama, S., Asanomi, Y., Boroevich, K. A., Sharma, A., Tsunoda, T., Matsukuma, K., Ichikawa, M., Sudo, H., Takizawa, S., Sakurai, T., Ozaki, K., Ochiya, T., & Niida, S. (2019). Risk prediction models for dementia constructed by supervised principal component analysis using miRNA expression data. *Communications Biology*, *2*(1), 1–8. https://doi.org/10.1038/s42003-019-0324-7

Sierra, J. (2018). Controlling Subjective Parameters of Convolution Reverb. *Proceedings of the 21st International Conference on Digital Audio Effects (DAFx-18)*. DAFx-18, Aveiro, Portugal.

Sony. (1999). *Sampling Digital Reverb: Operating Instructions (DRE-S777)*. https://pro.sony/s3/cms-static-content/operation-manual/3867715121.pdf

Soulodre, G. A., & Bradley, J. S. (1995). Subjective evaluation of new room acoustic measures. *The Journal of the Acoustical Society of America*, *98*(1), 294–301. https://doi.org/10.1121/1.413735

Stettner, A., & Greenberg, D. P. (1989). Computer graphics visualization for acoustic simulation. *Proceedings of the 16th Annual Conference on Computer Graphics and Interactive Techniques*, 195–206. https://doi.org/10.1145/74333.74353

Takahashi, D., Togawa, K., & Hotta, T. (2008). Objective measures for evaluating tonal balance of sound fields. *Acoustical Science and Technology*, *29*(1), 2–8.

Teret, E., Pastore, M. T., & Braasch, J. (2017). The influence of signal type on perceived reverberance. *The Journal of the Acoustical Society of America*, *141*(3), 1675–1682. https://doi.org/10.1121/1.4977748

Valimaki, V., Parker, J. D., Savioja, L., Smith, J. O., & Abel, J. S. (2012). Fifty Years of Artificial Reverberation. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(5), 1421–1448. https://doi.org/10.1109/TASL.2012.2189567

Valimaki, V., Parker, J. D., Savioja, L., Smith, J. O., & Abel, J. S. (2016). *More Than Fifty Years of Artificial Reverberation* [Paper presentation]. AES 60th International Conference, Leuven, BE.

van Dorp Schuitman, J. (2011). *Auditory Modeling for Assessing Room Acoustics* [Doctoral dissertation, Delft University of Technology]. https://repository.tudelft.nl/islandora/object/uuid%3A439c6688-e1c0-478e-b307-a61317c2b85b

van Dorp Schuitman, J., de Vries, D., & Lindau, A. (2013). Deriving content-specific measures of room acoustic perception using a binaural, nonlinear auditory model. *The Journal of the Acoustical Society of America*, *133*(3), 1572–1585.

Ware, C. (2021). *Information Visualization: Perception for Design* (4th ed.). Elsevier. https://doi.org/10.1016/B978-0-12-812875-6.00001-3

Weinzierl, S., Lepa, S., & Ackermann, D. (2018). A measuring instrument for the auditory perception of rooms: The Room Acoustical Quality Inventory (RAQI). *The Journal of the Acoustical Society of America*, *144*(3), 1245–1257. https://doi.org/10.1121/1.5051453

Wessel, D. L. (1979). Timbre Space as a Musical Control Structure. *Computer Music Journal*, *3*(2), 45–52. https://doi.org/10.2307/3680283

Wilkens, H. (1975). *Mehrdimensionale Beschreibung subjektiver Beurteilungen der Akustik von Konsertsälen* [Doctoral dissertation, Technical University of Berlin].

Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided Search: An Alternative to the Feature Integration Model for Visual Search. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(3), 419–433.

Wolfe, J. M., & Horowitz, T. S. (2004). What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience*, *5*(6), 495–501. https://doi.org/10.1038/nrn1411

Woszczyk, W., Begault, D. R., & Higbie, A. G. (2014). *Comparison and Contrast of Reverberation* [Paper presentation]. 137th Convention of the Audio Engineering Society, Los Angeles, CA.

Zahorik, P., Brungart, D. S., & Bronkhorst, A. W. (2005). Auditory distance perception in humans: A summary of past and present research. *ACTA Acustica United with Acustica*, *91*(3), 409–420.

8 APPENDICES

APPENDIX A: THE SPACEBUILDER IMPULSE RESPONSE LIBRARY

The Spacebuilder IR Library is vast collection of high-quality impulse responses intended for use in multichannel convolution reverb. Over a decade in the making, as of March 2022 it contains 11,107 multichannel IRs captured at multiple source and receiver positions in 206 venues. The Library uses an 80 second exponential sine sweep from 15 Hz to 47.5 kHz as an excitation signal, which is then deconvolved to produce impulse responses with high signal-to-noise ratios (Farina, 2000).

The Library is heterogenous with respect to the types of radiators used to excite venues and the types of microphones used to capture responses. Most commonly, however, the radiators used in measurement are an array of ten loudspeakers designed to roughly imitate the directivity of a musical instrument or small ensemble. Similarly, the most common capture instruments are an array of eight spaced microphones, four of which are omnidirectional and four of which are bidirectional. Generally, at each receiver position, measurements are made with the microphone array raised sequentially to three different heights (2m, 3m, and 4m). A more complete description of these loudspeaker and microphone arrays, and the motivations behind them, is given in Woszczyk et al. (2014).

IRs from the library were used to create stimuli for the experiments reported in Chapters 3-5. As the library was constantly expanding, and as the experiments occurred at different times, different sets of IRs were available when the experiments were conducted. At the time of the Chapter 3 experiment the library consisted of 7130 IRs; the IRs used were selected from this pool. At the time of the Chapter 4 and 5 experiments the library had grown to include 8154 IRs. This number represented measurements at 2718 source and receiver positions and at three microphone array heights (2m, 3m, 4m). These later chapters only made use of IRs captured at the 2m height, however, and so drew from only a third of the library: a pool of 2718 IRs.

Additionally, although the IRs in the library generally have eight channels, corresponding to the eight capture microphones, only the first two channels were used in the experiments. Sound stimuli were created, then, by convolving monophonic sound sources with two-channel

impulse responses, generally captured by two of the microphones in an eight-channel array. These two capture microphones were generally omnidirectional and spaced 2m apart.[36]

A visual summary of the IRs used to construct stimuli is shown in Figure 40. These 2581 distinct IRs represented 162 of the venues in the Library. The names of 12 representative venues also appear, arranged by the average measured T_{30} of the IRs belonging to them. The vertical axis shows a histogram of reverberation times (T_{30}) in the collection.

[36] Specifically, of the 2581 IRs used in experimental stimuli, 2416 used omnis spaced at 2m as capture microphones. As radiators, 1904 IRs used the ten-loudspeaker array mentioned above.

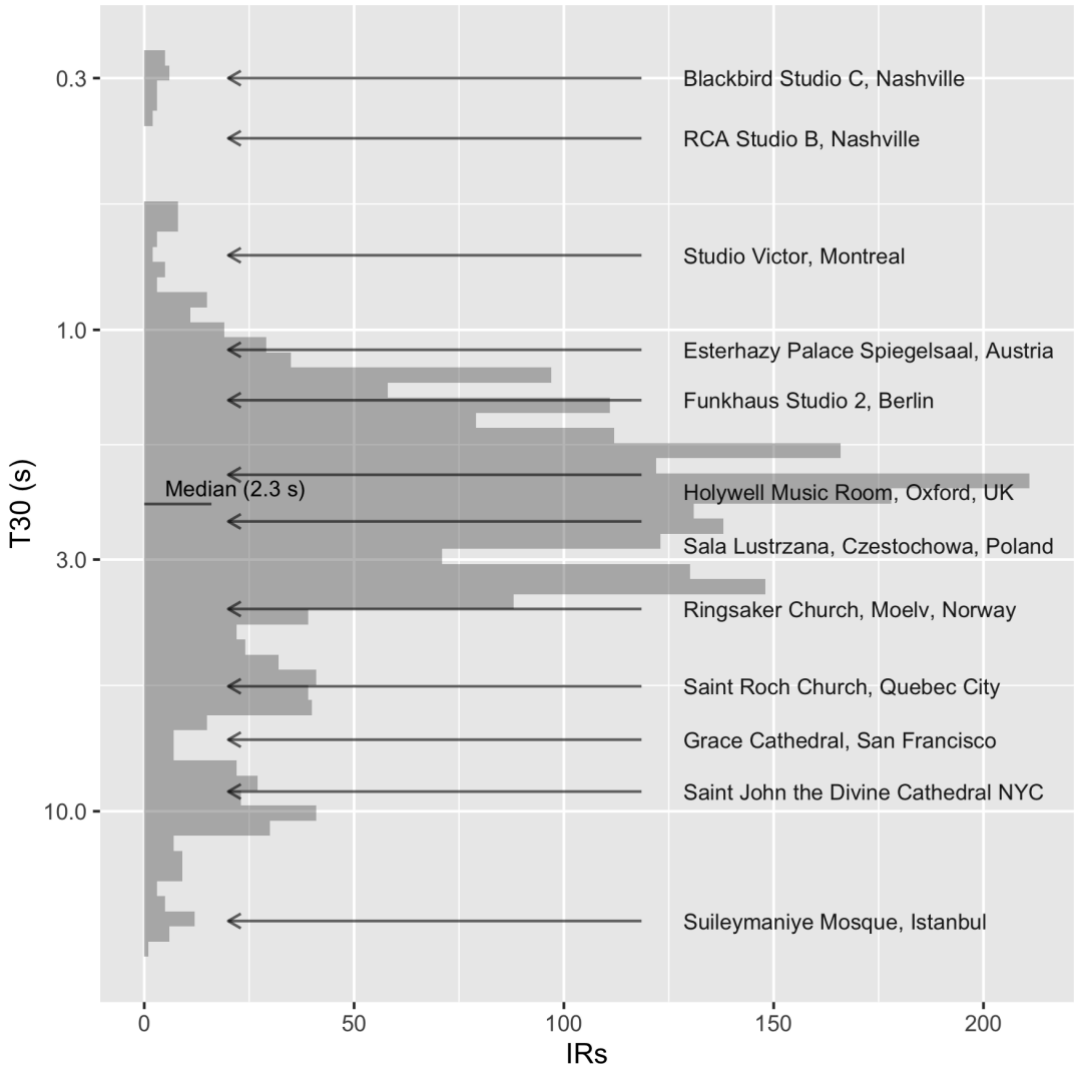


Figure 40 Summary of IRs used to create experimental stimuli

APPENDIX B: OBJECTIVE SIGNAL FEATURES

The experiments reported in Chapters 3-5 of this work all make frequent reference to *objective signal features*, defined here as physical measurements calculated from audio signals. These features fall into two broad categories. Those in the first category are calculated from impulse responses and are referred to as *IR features*, while those in the second category are calculated from the outputs of convolutions between IRs and sound source signals. These convolution outputs are called "wet" signals in this work, and features in this category are referred to as *wet features*.

Most of the work's signal features are described in detail in the main text. This appendix serves to define those that are not explained elsewhere and to give additional information about some that are. In this appendix, features are addressed in roughly the same order that they appear in Table 4. One additional feature that does not appear in this table, but which is referenced in Chapters 4 and 5, the *IR spectral slope*, is discussed at the end of this section.

EDT, T_{30}, C_{80}, *IACC*

Signal analysis for this research was performed mostly in MATLAB. Specifically, the *AcMus - Room Acoustic Parameters* toolbox (Masiero, 2007) was used to calculate the four ISO 3382 signal features used in Chapters 3-5: *EDT*, T_{30}, C_{80}, and *IACC*.

Regarding *IACC*, the feature was computed on the left and right channels of a stereo impulse response, generally captured via spaced omni-directional microphones, rather than on the left and right ear outputs of a dummy-head microphone as is specified in ISO 3382. These stereo IRs were used to generate stimuli that were always auditioned over headphones, however, meaning that the *IACC* was related to signals that were auditioned binaurally, even if they were not measured with a binaural capture device.

Two version of the *IACC* were computed, one on the first 80 ms of the IR, and one on the remainder of the IR, after 80 ms. The first measure is referred to simply as *IACC* in the text, while the second is referred to as $IACC_{late}$. As suggested by Okano et al. (1998), the value reported is the average from the 500, 1k and 2k Hz octave bands.

In general, all features that examined temporal regions of IRs considered the IR to begin at its "onset". The onset was defined as the earliest time at which the squared IR signal rose above 1/100 of its maximum value. That is, features such as the C_{80} and *IACC,* which examined the "first" 80 ms of the IR, considered the region between the onset and 80 ms after the onset.

Number of peaks, NED mixing time, degree of curvature

To compute these features, MATLAB implementations written the author were used. The *number of peaks* and *NED mixing time* features were discussed in Chapter 2. The *degree of curvature* is drawn from ISO 3382-2 (2008) and is essentially a ratio of reverberation time estimates over different regions of the energy decay curve, as specified by Equation 18.

$$C = 100 \times \left(\frac{T_{30}}{T_{20}} - 1\right)$$

Equation 18 *Degree of Curvature*

pRev, pClar, pASW, pLEV

These features, employing a binaural model and discussed in section 2.2.2.4, were computed using a closed-source implementation provided by their creator, Jasper van Dorp Schuitman. It should be noted that these features are non-linear and are designed to operate on loudness calibrated signals, where the precise relationship between signal level and SPL at the listener's ears is known. In this work, signals were not calibrated. Rather, sound stimuli were normalized to -18 LUFS and listening test subjects were invited to adjust playback levels as they wished, even though the features were designed on the assumption that a full-scale signal would correspond precisely to a playback SPL of 92 dB. Although these features had good predictive performance in the given experimental setup, in the subjective opinion of the author, their predictions would likely have been yet more accurate had playback levels been properly calibrated.

It is also important to note that, similarly to the *IACC* features discussed above, these four features were originally designed to analyze binaural signals (e.g., signals captured by a dummy head microphone). They were not applied to such signals in this work. Rather, Chapter 3 applied them to synthetic reverberation intended for loudspeaker reproduction, and chapters 4 and 5 applied them to reverberation measured by widely spaced microphone pairs. These latter types of signals differ greatly from binaural signals in nature of their inter-channel timing and level differences. In the case of binaural signals, for instance, inter-channel time delays (ITDs) are limited to small values due to the close spacing of the ears. Widely spaced microphone signals exhibit much higher limits on maximum ITD, while synthetic reverberation signals, in principle, have no ITD constraints.

Given these variations in inter-channel differences, it bears considering whether a binaural model such as Schuitman's, trained on binaural recordings, could still produce perceptually accurate predictions while analyzing other types of signals. For the sake of argument, consider a simple sound field consisting of a sound source at exactly 1 degree off-axis. A dummy head microphone exposed to this sound field, with an inter-capsule spacing of 17.5 cm, would capture two signals with an ITD of about 0.009 ms.[37] Next, consider a very simple binaural model, one that predicts sound source angle solely using ITD. This simple binaural model, when trained on such signals, would learn to predict a 1 degree source angle from this 0.009 ms ITD.

[37] This discussion relies on the simplifying assumptions of an infinitely distant sound source, an acoustically transparent dummy head, and a speed of sound in air of 344 m/s. Under such assumptions, the ITD captured by a pair of microphones can be approximated by $\frac{a}{c}\sin\theta$, where θ is sound source angle from the plane separating the microphones, a is the microphone spacing, and c is the speed of sound.

A widely spaced microphone pair would respond differently to this sound field. At a spacing of 2 m, such microphones would capture signals with an ITD of 0.1 ms. If these spaced microphone signals were analyzed by the simple binaural model discussed above, the model would predict a much larger source angle of 12 degrees. In a sense, this prediction would be incorrect, as it would overestimate the physical angle of the source. This 12 degree prediction might still be perceptually correct, however, as it might correspond closely to the sound source angle perceived when these spaced microphone signals were auditioned over headphones. In other words, headphone auditioned spaced microphone signals give rise to spatial auditory imagery, just as binaural signals do, and aspects of this imagery could conceivably be predicted by a binaural model, even a model not trained on such signals.

Apropos this research is the question of whether a more complex binaural model would also succeed in correctly predicting the spatial imagery associated with spaced microphone signals. In the case of Schuitman's model, no studies have yet formally tested its accuracy with such signals, so the author performed a series of informal tests. While conducting this research, the author spent many hours auditioning experimental stimuli, subjectively evaluating their ISO attributes, and comparing these subject evaluations with the predictions of *pRev, pClar, pASW* and *pLEV*. In the opinion of the author, these features performed well, and succeeded in making perceptually accurate predictions on the non-binaural signals used in this work. Although more research is needed to confirm these impressions, the author's subjective experience can perhaps be considered circumstantial evidence that the features are indeed reasonably effective for predicting reverberation attributes in at least some non-binaural contexts.

Log attack time, spectral slope (wet), spectral skew, spectral decrease, spectral flatness

These features were computed using the Timbre Toolbox (Peeters et al., 2011). *Log attack time* was calculated on the IR only; the other features were calculated on wet signals. The "ERB fft" auditory model was used; all other analysis parameters were left at their defaults, including the window hop size (5.8 ms). The median value over all signal windows was taken as the final value of each feature.

B.1 Spectral slope (IR)

As discussed in section 4.2.2, the *spectral slope (IR)* was defined as the slope of the line of best fit through an IR's smoothed magnitude spectrum. Smoothing was performed via a moving average over the frequency axis, using windows 1/3 octave wide and with a hop size of 1/12 octave. Frequencies from 63 Hz to 12500 Hz were considered, and only the first 300 ms of the IR was analyzed. A MATLAB implementation by the author was used, which is given below.

```
function [ SSIR ] = SpectralSlopeIR( ir, fs )
% Inspired by "Deviation of Level" from Takahashi et al, 2008

  % ir should a contain single-channel IR starting at its onset
  ir300 = ir(1:ceil(fs*0.3)); % only keep first 300ms

  % get dB magnitude spectrum
  irFFT=fft(ir300);
  Lo2=round(length(irFFT)/2); % fft length/2
  freqVec=((0:Lo2-1)/Lo2)*fs/2; % magnitude spectrum bin frequencies
  irMagSpec=20*log10(abs(irFFT(1:Lo2)));

  % vector of exponents to make center frequencies of 1/12 oct bands
  % from two bands below 63 Hz (1000*2^(-4-(2/12)))
  % to two above max freq. of 12.5 kHz
  % (1000*2^(log2(12500/1000)+2/12))
  maxFreq = 12500;
  exponents = [(-4-(2/12)):(1/12):(log2(maxFreq/1000) +(2/12))];

  % vector of 1/12 octave band center frequencies
  cfs12 = 1000*2.^exponents;

  % smooth spectrum by averaging over a 1/3 oct. sliding window
  smoothedMagSpec = zeros(length(cfs12)-4,1);
  for b = 3:(length(cfs12)-2)
    % lo and high band edges for 1/3 octave smoothing
    lo = cfs12(b-2);
    hi = cfs12(b+2);
    selFreqIds = find(freqVec>lo&freqVec<hi);
    smoothedMagSpec(b-2) = mean(irMagSpec(selFreqIds));
  end

  smoothedCfs = cfs12(3:end-2);  % bin freqs of smoothed spectrum

  % regress smoothed magnitude spectrum onto log of frequency
  y = smoothedMagSpec;
  X = [ones(1,length(smoothedCfs)); log(smoothedCfs)]';
  b = X\y;
  y_hat = X*b;

  SSIR = b(2);  % spectral slope is regression coefficient b1

  % draw a diagnostic plot
  if 1
    hold on
    plot(freqVec,irMagSpec,'LineStyle','-','Color',[0 0 1 0.1])
    plot(smoothedCfs,smoothedMagSpec,...
      'LineStyle','none','marker','o','markersize',6);
    plot(smoothedCfs,y_hat,'LineStyle','-','Color',[1 0 0],...
   'LineWidth',2);
    hold off

    axis([smoothedCfs(1) smoothedCfs(end) -Inf Inf ]);
    set(gca,'Xscale','log');
    title(['IR Spectral Slope: ' num2str(SSIR)]);
  end
end
```

APPENDIX C: SOUND SOURCES

A total of four different monophonic sound sources were used in the three experiments reported is this work. All four appeared in Chapter 3 (orchestra, chorus, drums, jazz voice), while two appeared in Chapter 4 (drums, jazz voice). Chapter 5 only made use of the drums sound source. In this appendix each sound source will briefly be discussed.

The drums, jazz voice and chorus sources (Figure 41, Figure 42, Figure 43) were all recorded at McGill University in dry, but not anechoic, conditions. Drums consisted of a standard drumkit playing a rock pattern. Jazz voice is a recording of solo female vocalist singing the first verse of Sarah Vaughan's *Day by Day*, while chorus is recording of a vocal quartet singing renaissance music. Spectrograms were generated using the Sonic Visualizer software (Cannam et al., 2010).

The orchestra source (Figure 44) was created by summing individually recorded orchestral instruments playing an excerpt of romantic orchestral repertoire (Bruckner's Symphony no. 8, 2nd movement). These anechoic recordings, made at the Helsinki University of Technology, were intended to support research in concert hall acoustics (Pätynen et al., 2008).

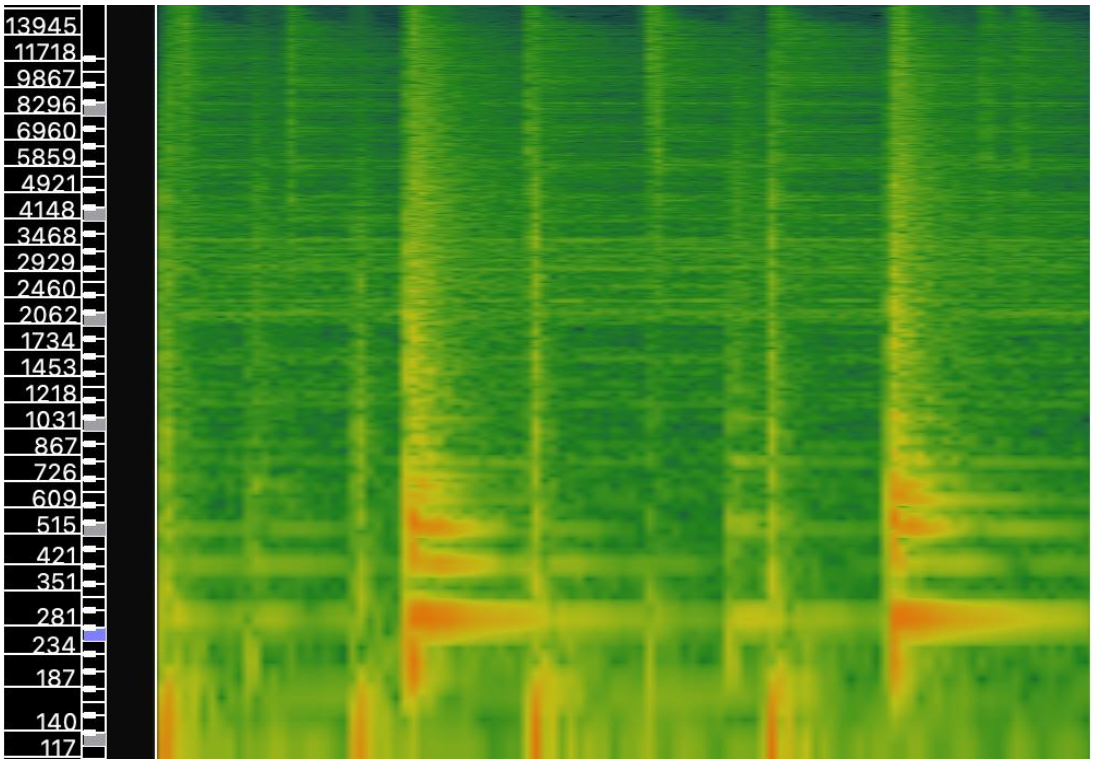


Figure 41 Drums source (2.3 s)

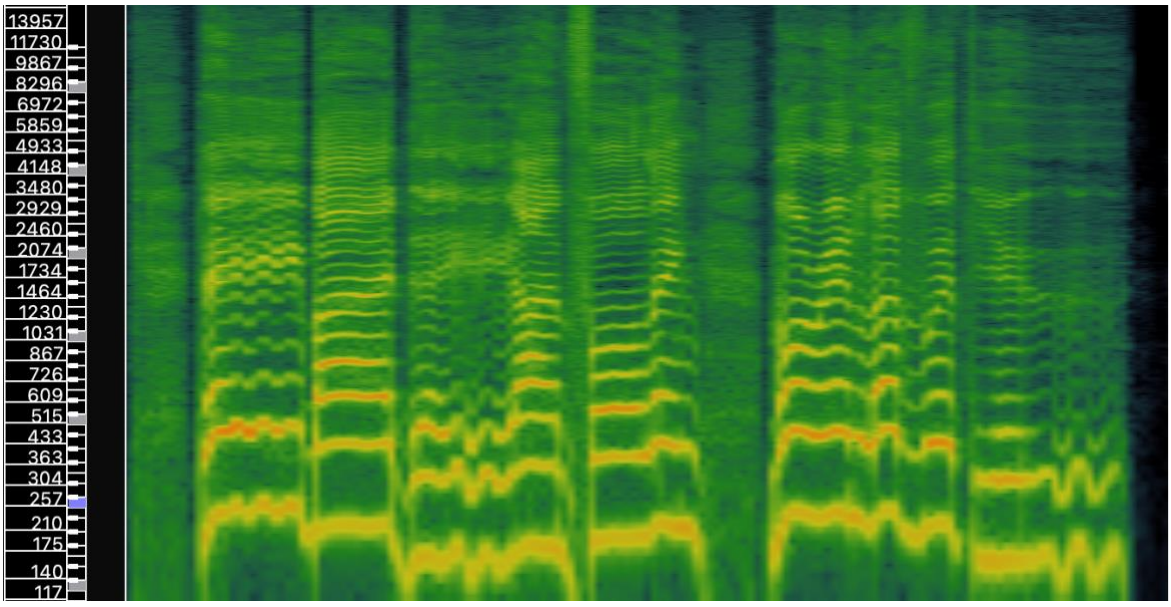


Figure 42 Jazz voice sound source (6.5 s)

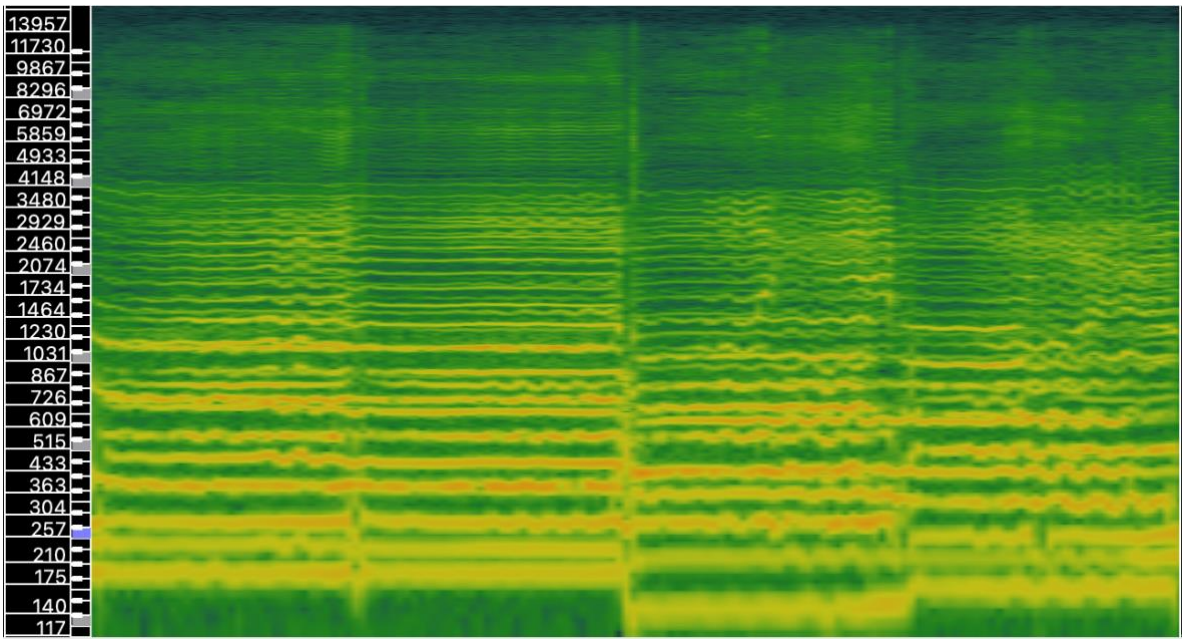


Figure 43 Chorus sound source (6.5 s)

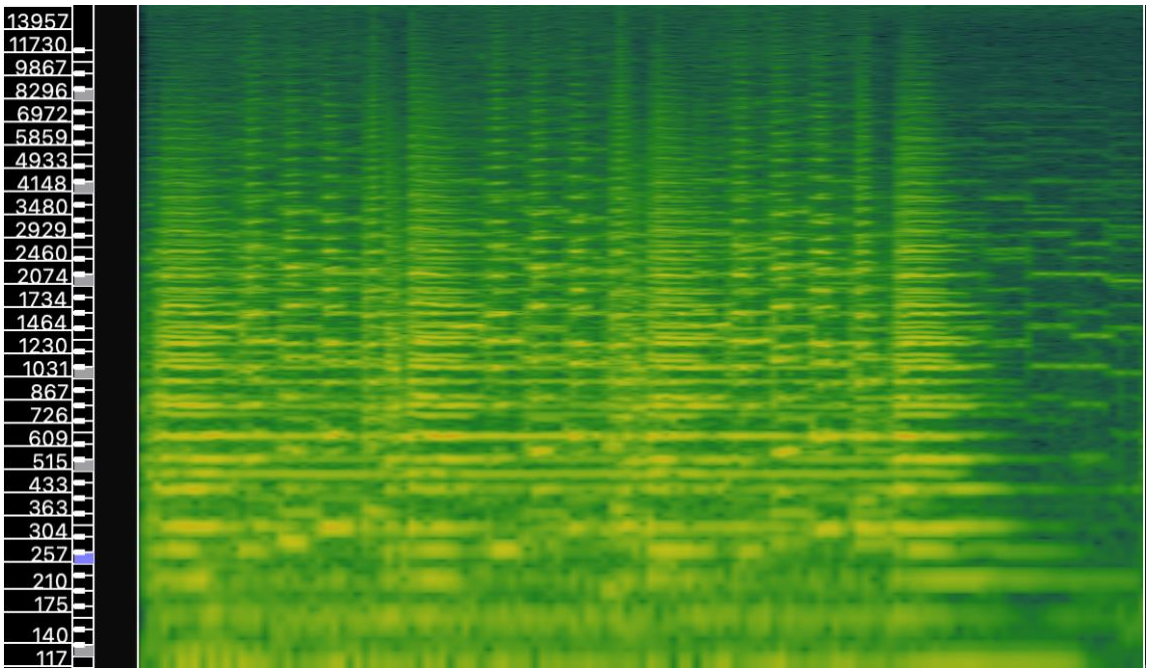


Figure 44 Orchestra sound source (5.5 s)

APPENDIX D: STATISTICAL TECHNIQUES

This appendix describes some of the statistical tools used in Chapter 3 of the dissertation. The first half of Chapter 3 was concerned with building predictive models of algorithmic reverb preset labels. A technique called Supervised Principal Components Analysis Logistic Regression (SPCA-LR) was used for this purpose. SPCA-LR is described first. Significance tests were then conducted on the resulting models to test whether the SPCA-LR algorithm had succeeded in capturing meaningful relationships between labels and signal features. These model significance tests are explained in the second section. The third section of this appendix describes the t-tests used to investigate relationships between objective signal features and the textual responses given by subjects in the subjective experiment reported in the second half of chapter 3.

D.1 Supervised Principal Components Analysis

Section 3.1 of this work investigated statistical relationships between a set of 36 objective signal features and 12 descriptive labels in a collection of 702 algorithmic reverb presets. Specifically, it built predictive models that accepted signal features as input, and output the probability of a label being present on a preset IR. The chapter mentioned logistic regression as the basic architecture of these predictive models.

A difficulty with using logistic regression in this context, however, is that standard logistic regression only uses a single predictor variable. Although separate logistic regression models could have been built and compared for each label-signal feature combination, such models would not have been able capture complex relationships between the variables. For example, a single-variable logistic regression model might give sub-optimal predictions in cases where a label was best characterized by, say, both a low late treble ratio and a high *IACC*.

One modeling approach that can capture multi-variable relationships, and which could theoretically have been used, is multiple logistic regression. Multiple logistic regression extends basic logistic regression by adding new terms for each input variable. An example with p input variables is shown in Equation 19.

$$f(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}}$$

Equation 19 Multiple logistic regression

Unfortunately, however, our data set was too small to use this standard technique. A well-known statistical rule of thumb states that in multiple logistic regression, at least 10 positive examples of a class are required for each predictor variable (Peduzzi et al., 1996). In a dataset with 36 predictor variables, then, complying with this "rule of ten" would require at least 360 examples of each label. This requirement was not met by our data, which contained only 137 examples of the most frequently occurring label ("plate") and 35 examples of the least frequently occurring label ("dark"). Violating this rule increases the chances statistical "overfitting", or of producing models that predict the training data well, but which generalize poorly to novel data. Overfit models would have been unlikely to correspond to dimensions of perceptual variation in natural IR libraries.

Supervised Principal Components Analysis (SPCA) is a regression technique specifically designed to guard against overfitting in high-dimensional settings such as ours, while also being capable of capturing complex relationships between variables (Bair et al., 2006). SPCA models are built with the three-stage algorithm explained below. The stages consist of feature selection, dimensionality reduction, and model fitting.

Stage 1: feature selection

In the first stage, a subset of the 36 features is selected for inclusion in the model. Prior to selection, the features are ranked by the strength of their association with the given label. In this work, association strength is measured using Spearman correlation. Figure 45 shows the 36 features ranked by their strength of association with the label "dense".

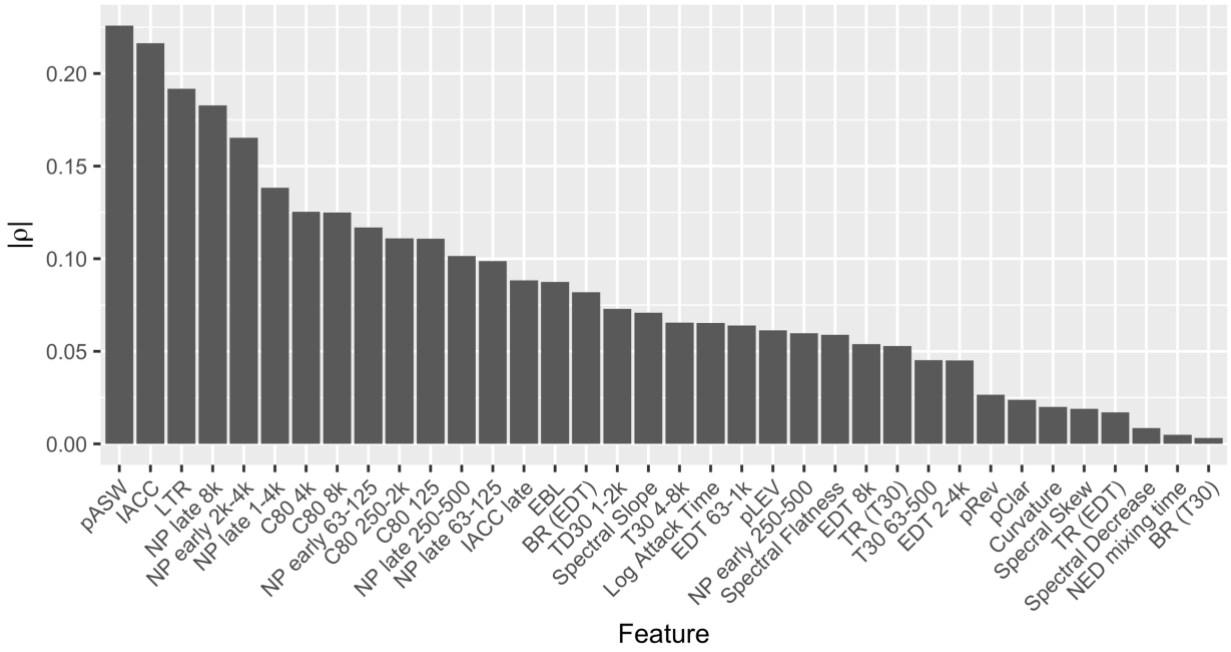


Figure 45 Signal features ranked by strength of association with label "dense"

Following ranking, the top M features are retained, and the bottom $36 - M$ features are discarded. The number of retained features, M, is chosen using cross validation, as explained below.

Stage 2: dimensionality reduction

In the second stage, the *M* retained features are submitted to a Principal Components Analysis (PCA). Unlike a standard PCA, which would be performed on all input variables, this "supervised" version contains only those variables most strongly associated with the outcome variable (the label). Unrelated or “noisy” features are excluded.

Figure 46 shows the first two principal components of a PCA on the four features most strongly associated with "dense". "Dense" presets are shown in blue; presets without the "dense" label are shown in red. Note that most "dense" presets have a high score on the first PC (i.e., they're far to the right on the x-axis). This first PC, which itself is a linear combination of four features, does a good, if imperfect, job of discriminating those presets with the label from those without it.

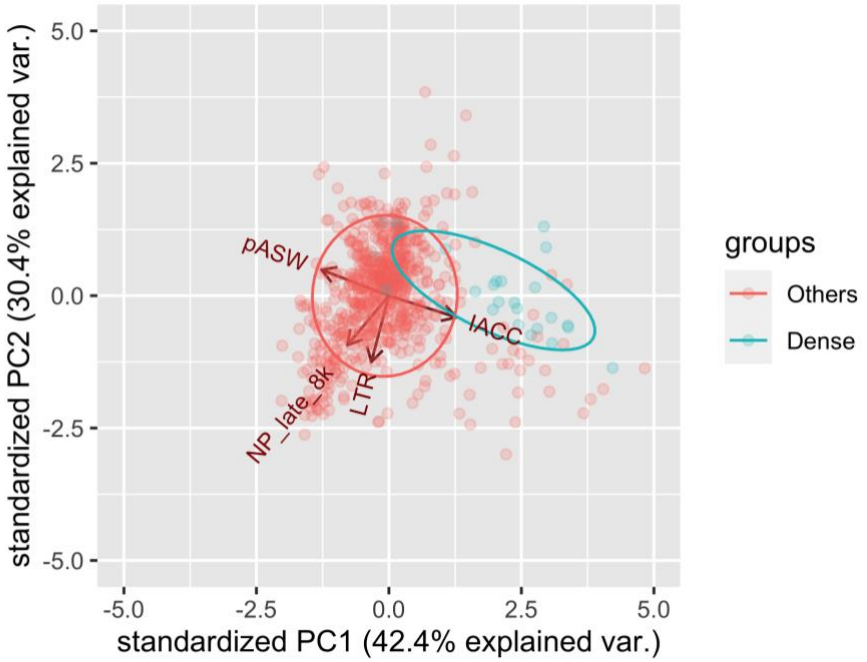


Figure 46 Biplot of PCA on features most strongly associated with *dense*

Stage 3: model fitting

In the third stage, a standard logistic regression model is fit to the label and the first PC. As discussed earlier, this model has the form of Equation 20 where $\hat{y}$ is the predicted probability of the label being present on the IR and $\hat{\beta}_0$ and $\hat{\beta}_1$ are parameters fit by maximum likelihood (Friedman et al., 2009). As specified in Equation 20, t, the first principal component, is a weighted combination of *M* signal features, with the weight vector $w_1, w_2, \ldots w_M$ determined through PCA.

$$\hat{y} = \frac{1}{1 + e^{-(\hat{\beta}_0 + \hat{\beta}_1 t)}}$$

Equation 20 Logistic regression

$$t = w_1 x_1 + w_2 x_2 + \cdots + w_M x_M$$

Equation 21 First principal component

In the example above, the four signal features that contribute to t are *pASW, IACC,* TR_{late} and *NP late 8k*.

Cross-validation to choose hyperparameter M

As mentioned above, the *SPCA* algorithm contains one free variable, or "hyperparameter", *M*, which determines how many of the ranked features to include in the model. Setting *M* too low may exclude useful variables, while setting it too high may result in too many "noisy" variables being included, which can also hurt performance. In this work, as in similar applications (e.g., Shigemizu et al., 2019), *M* is chosen using cross-validation.

In cross-validation, the dataset is divided into partitions, some of which are used for model building, or "training", and some of which are used for model evaluation, or "testing". Ten partitions were used in this work. In each of ten iterations, one partition was assigned the role of "training" data, and the other nine partition were combined and assigned the role of "test" data. Multiple SPCA models were built on the training data, one for each value of M, from 1 to 36. Each of these models was then evaluated on the test data, using a metric called the AUC.[38] Ten iterations resulted in ten performance estimates for each M, which were then averaged. Figure 47 shows examples of the average AUC plotted against M for the "dense" label model. For this label, optimal performance was seen with M equal to four.

Once an optimal M was chosen through cross validation, all partitions were then recombined, and a "final" model was built from the entire dataset using this optimal M. These "final" models, which consist of weights $w_1, w_2, \ldots, w_M$ for M features and parameter estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, are reported in Figure 10. Note that the four features used in the *dense* model are the same as those shown in Figure 46, above.

[38] The AUC, or Area Under the receiver operating characteristic Curve, is a metric for evaluating classification algorithm performance (A. P. Bradley, 1997). When a classifier performs no better than chance it takes on values near 0.5; when a classifier performs perfectly it takes on a value of 1.

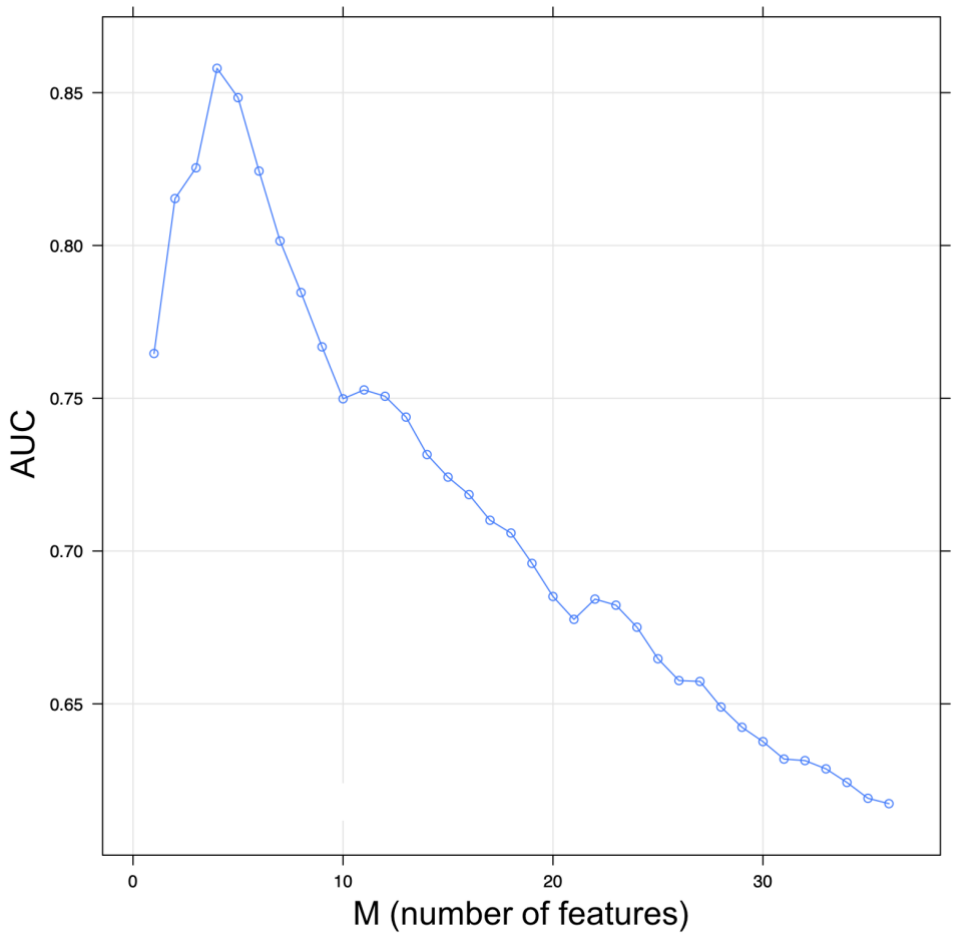


Figure 47 Estimated performance of "dense" label models by value of M

D.2 Randomization tests to calculate model p-values

One key question about any regression model concerns the strength of the relationship between predictor and outcome variables, and whether this relationship is stronger than would be expected by chance. Typically, this question is answered through some form of hypothesis test, where the observed value of a statistic is compared with its theoretical distribution under a null hypothesis. In standard logistic regression, for example, the *z*-statistic associated with parameter β_1 follows a normal distribution under the null (Field et al., 2012). Comparing an observed *z*-statistic with its likelihood under the null produces a *p*-value, which can be interpreted as the probability that a relationship of the observed strength would occur by chance. In other words, a low *p*-value suggest that a meaningful relationship exists between the predictor and outcome variables.

The question of whether meaningful relationships exist between predictors and outcome variables, that is, whether meaningful relationships exist between features and labels, is also of interested in this work. Unfortunately, though, the feature selection stage in the SPCA algorithm, described above, changes the distribution of the β_1 parameter's *z*-statistic under the null, rendering it non-normal. This makes the significance test above difficult to apply. An alternative method is needed to calculate *p*-values for our models.

In this work, a non-parametric method called a randomization test was used to evaluate models (Legendre & Legendre, 2012). In a randomization test, instead of comparing a test statistic with its theoretical distribution under the null, the statistic is compared instead with its empirical distribution under the null. That is, multiple datasets are generated for which the null hypothesis is known to be true, and the test statistic is calculated for each of these artificial datasets. The calculated test statistics then constitute an empirical null distribution, with which the observed test statistic can be compared. Randomization tests are flexible in the choice of test statistic; this work uses the AUC, the same metric used to set hyperparameter M in the SPCA discussion above.[39]

To calculate *p*-values for our models then, distributions of the AUC under the null hypothesis were required. Calculating these AUC distributions, in turn, required many datasets for which the null hypothesis was known to be true. To construct these null distributions, 2000 new datasets were created in which any existing relationship between features and label was deliberately destroyed. These "noise" datasets were produced by stripping the labels from each preset IR and randomly reassigning them to other IRs in the collection.[40] For each these noise datasets, SPCA-LR models were then constructed using the method described earlier, in section D.1. This produced 2000 AUC values for each label, created from datasets in which the null hypothesis was known to be true. These 2000 values formed an empirical distribution of the AUC under the null.

[39] All references to AUC in this section refer to the average AUC calculated via ten-fold cross validation. This value might more properly be called an "out-of-sample AUC estimate", as it uses cross validation to estimate the performance of the model, as measured by the AUC, on unseen data. The term "AUC" is used for simplicity.

[40] More concisely, the "noise" datasets were constructed via "permutations" of the outcome variable vector. For this reason, randomization tests are sometimes called "permutation tests".

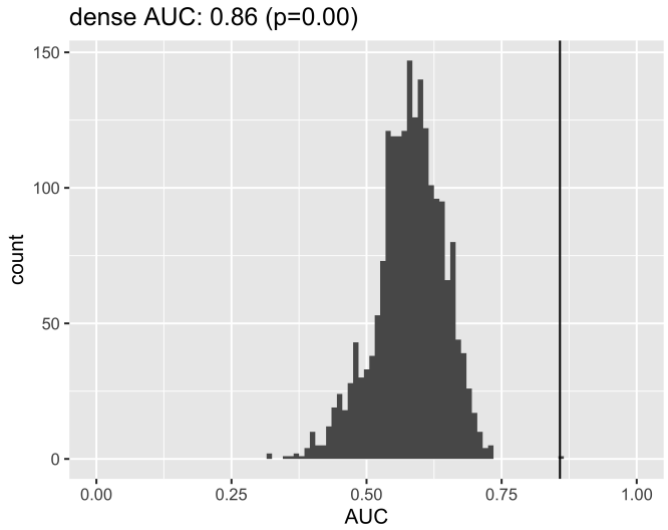


Figure 48 Randomization test results for *dense*

From these empirical null distributions, *p*-values for the observed AUCs were computed. Two examples of these computations shown in Figure 48 and Figure 49. Figure 48, for example, shows the distribution under the null hypothesis of the AUC for the "dense" label model. The vast majority of null hypothesis AUCs are under 0.75, while the AUC observed on the genuine data, shown by the vertical line, is 0.86. As only one null hypothesis AUC out of 2000 is above 0.86, this corresponds to a *p*-value of 1/2000 (or rounded to two decimal places, 0.00). This suggests that the SPCA-LR algorithm found a relationship between the features and the label "dense" that was much stronger than would be expected by chance.

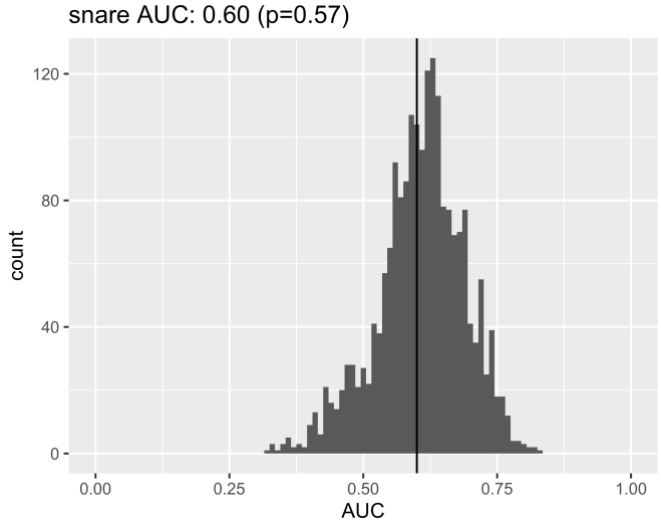


Figure 49 Randomization test results for *snare*

Conversely, Figure 49 shows the null distribution of the AUC for the *snare* label model. Again, the AUC observed on the genuine data, 0.6, is shown by a vertical line. This is near the middle of the null distribution, and well within the range of values that would be expected if the "snare" label had no characteristic signature in the feature set. The high *p*-value of 0.6 suggests that the SPCA-LR algorithm was unable to find any meaningful relationship between the features and the "snare" label.

Observed AUCs for each of the twelve models, along with their associated *p*-values, are shown in Table 13.

Table 13 Observed AUCs and associated *p*-values for the 12 models

	AUC	p
plate	0.68	0.00
dense	0.86	0.00
ambience	0.78	0.00
dark	0.88	0.00
chamber	0.68	0.00
bright	0.72	0.01
vocal	0.67	0.01
rich	0.71	0.01
drum	0.69	0.04
snare	0.60	0.57
space	0.57	0.62
percussion	0.51	0.86

D.3 T-tests to explore feature and attribute term associations

In section 3.2.2.3, *t*-tests were used to explore the relationships between signal features and attribute terms in the experimental data. This section explains how these tests were carried out.

As mentioned in the main text, the triadic comparisons in the experimental trials were first decomposed into sets of two pairwise comparisons. Within a trial, if stimulus 1 was selected as different and described as the "brightest" of the three, this was taken to mean that it had more brightness than stimulus 2 (pairwise comparison one) and also more brightness than stimulus 3 (pairwise comparison two). This one triadic comparison, then, produced two relative judgements.

A statistical method was then needed to see which signal features were most strongly associated with these relative judgements. If it were found, say, that stimulus pairs

differing in brightness also consistently differed in *early bass level* (*EBL*), this would constitute evidence of an association between *EBL* and brightness. On the other hand, if the average difference in *EBL* within these pairs was small, or zero, this would be evidence for a lack of association.

To investigate these trends empirically, single sample *t*-tests on feature differences were used. A single sample *t*-test produces a *p*-value and a *t*-statistic. The *t*-statistic, defined as the sample mean divided by its standard error, can be considered a measure of "how different from zero" the observed mean is. Larger *t*-statistics suggest stronger associations between feature and attribute. The *p*-value is the probability of observing the given *t*-statistic under the null hypothesis that the true mean of the data is zero. Figure 50 and Figure 51 show graphically how these tests were carried out.

Figure 50, for example, examines the association between the C_{80} *4k* feature and the attribute of sound source distance. In the large table on the bottom left, each row represents one pairwise comparison. Within each pair, one stimulus was judged to be "closer" and one was judged to be "farther". The difference in C_{80} *4k* within each pair was then calculated by subtracting the C_{80} *4k* of the farther stimulus from the C_{80} *4k* of the closer one. These differences are shown in the bottom right panel. The top right panel shows a histogram of the differences.

A *t*-test was then performed on the difference data to examine the strength of association between the attribute and feature. In this case, the histogram leans heavily to the left, suggesting that C_{80} *4k* differences were indeed associated with terms related sound source distance. This intuition is confirmed by the *t*-test results, shown on the top left, which include a *t*-statistic large in magnitude (-13.36) and a small *p*-value.

A second example is shown in Figure 51. This figure examines the association between the C_{80} *4k* feature and the terms related to brightness. Here, the histogram on the top right shows a distribution whose mean seems close to zero, suggesting a lack of association between C_{80} *4k* and brightness. This intuition is confirmed by the *t*-test, whose *t*-statistic (-0.725) is much smaller in magnitude than in the previous example. Additionally, the high *p*-value (0.47) gives no evidence to reject the null hypothesis that the true mean of differences is zero.

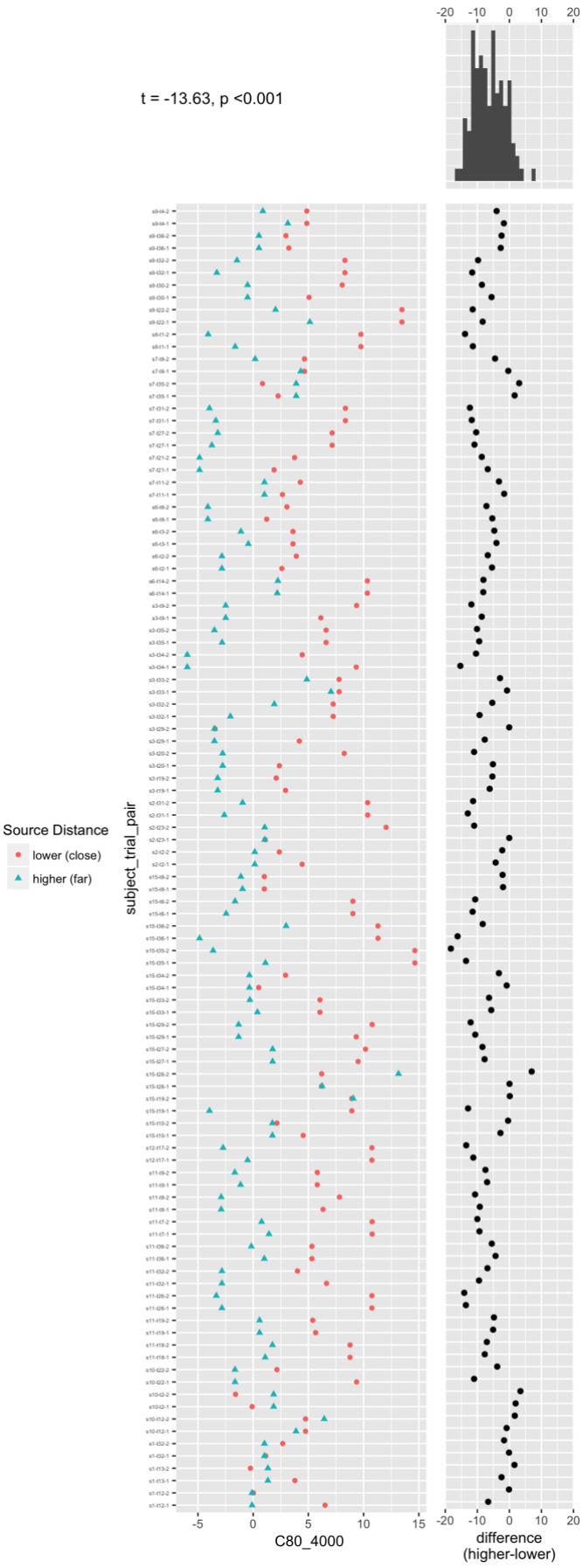


Figure 50 *T*-test on association between distance terms and *C80 4k*

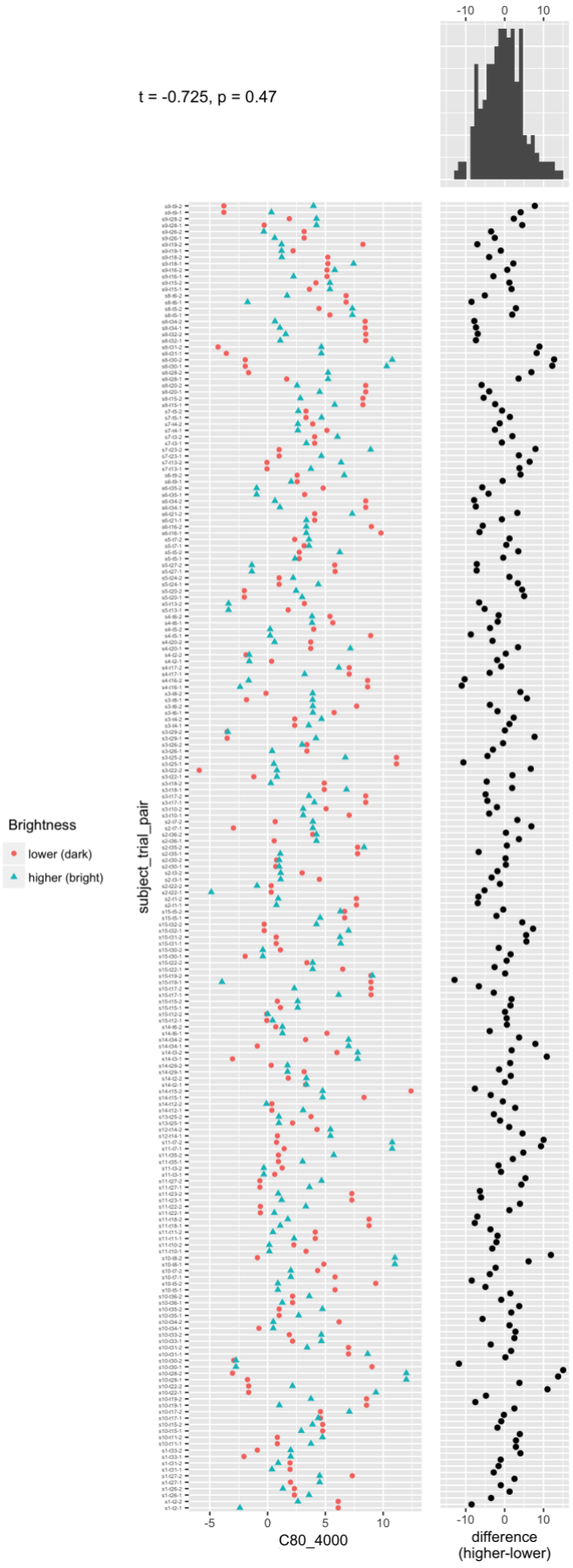


Figure 51 *T*-test on association between brightness terms and *C80 4k*

APPENDIX E: IR SAMPLING ALGORITHM FOR CHAPTER THREE EXPERIMENT

The experiment reported in section 3.2 required subjects to compare groups of three stimuli. As mentioned in the main text, the stimulus triplets were chosen such that they exhibited large variations in label model probabilities, but minimal variations in ISO 3382 model attributes. The rationale for minimizing variation in ISO attributes was to focus subjects' attention on other sources of perceptual variation, not present in the ISO model, that might exist within the stimuli. This section lays out the algorithm by which stimulus triplets were assembled.

The stimulus selection algorithm consisted of six steps which were run in a loop. Each iteration through the loop produced one triplet of stimuli, suitable for one trial of the experiment. In each of the 36 experimental conditions (four sound sources by nine label models), the loop was run 15 times, to produce 15 unique triplets for the 15 subjects.

The six steps were as follows.

Step one: create groups of "low" and "high" probability stimuli

Initially, a very large set of potential stimuli was created by convolving the four sound sources with each IR in the Spacebuilder library. For each of these potential stimuli, label probabilities were calculated using one of the nine label models under investigation. As one goal of the experiment was to test whether extreme values of modeled label probabilities could be discriminated, stimuli with label probabilities close to the median were discarded, and only those with extreme values were retained. Specifically, all stimuli with label probabilities between the 5th and 95th percentile were discarded. Of the remaining stimuli, those with label probabilities below the 5^{th} percentile were placed in a “low” group, and those with probabilities above the 95^{th} percentile were placed in a “high” group.

Step two: randomly select a trial structure

In the second step of the algorithm, a trial structure was chosen at random. The structure indicated the arrangement of low and high probability stimuli within the trial. Trials with "low-high-high" structure contained one low probability stimulus and two high probability stimuli, whereas trials with low-low-high" structure contained one low probability stimulus and two high probability stimuli.

Step three: select 100 random "probe" points in the space of ISO attributes

Third, a set of 100 random "probe" points were chosen in the four-dimensional space of the ISO attributes. ISO attribute values were estimated in this case by the binaural model discussed in section 2.2.2.4, so, more precisely, this step selected 100 "probe" points in the four-dimensional space defined by features *pRev, pClar, pASW* and *pLEV*.[41] Each point corresponded to a distinct set of attribute values. For example, one probe point might designate dry, clear, narrow, unenveloping reverberation, while another might designate wet, unclear and wide-sounding, enveloping reverberation.

Step four: create 100 "probe triplets"

Next, for each probe point, the algorithm attempted to find a group of three stimuli that varied in their label probabilities while also being close to the probe point. This corresponded to searching for three stimuli that had approximately the same amount of reverberance, clarity, source width and envelopment, but had wildly different probabilities of a label such as *vocal*. A group of three stimuli having these properties was known as a "probe triplet".

Specifically, probe triplets were constructed by selecting the three stimuli closest to each probe point, subject to two constraints. First, depending on the trial structure, the three points needed to include either two "low" stimuli and one "high" stimulus, or one "low" stimulus and two "high" stimuli. Second, the three stimuli were required to use

[41] To ensure that these 100 probe points lay in regions of the feature space that were relative densely populated by stimuli, the points were actually selected in a 2D PCA projection and then rotated back into the original four-dimensional feature space.

IRs measured in three different venues. The second constraint existed to help ensure that the three stimuli were perceptually distinct. Without it, the algorithm tended to choose two of the three stimuli from nearby positions in the same venue, resulting in two stimuli being nearly perceptually identical.

Step five: select the probe trio with the smallest amount of variation in ISO attributes

Step four yielded 100 potential stimulus triplets. These triplets, however, were unequal in the homogeneity of ISO attribute features. Some trios were clustered closely around their probe point, indicating similar ISO attribute feature values, while others were more spread out, indicating more variation in ISO attributes. In this step, the 100 probe trios were ranked by their spread in the ISO attribute feature space. Triplet spread was quantified using average Euclidean between the three stimuli and the probe point. From the 100 candidates, the single trio with the smallest spread, and hence the greatest similarity in ISO attribute features, was selected as a stimulus triplet to be used in the experiment.

Step six: remove the three stimuli in the chosen triplet from further consideration

Lastly, to ensure that no stimuli were repeated between subjects within an experimental condition, the three stimuli in the triplet were removed from future consideration. Steps one to five were then repeated again, once for each subject in the experiment.