U.PORTO

MESTRADO
MULTIMÉDIA - ESPECIALIZAÇÃO EM MÚSICA INTERATIVA E DESIGN DE SOM

SOUNDSCAPE GENERATION USING WEB AUDIO ARCHIVES

PAULO TEIXEIRA

M
2019

FACULDADES PARTICIPANTES:

FACULDADE DE ENGENHARIA
FACULDADE DE BELAS ARTES
FACULDADE DE CIÊNCIAS
FACULDADE DE ECONOMIA
FACULDADE DE LETRAS

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Soundscape Generation Using Web Audio Archives

Paulo Jorge Fernandes Teixeira

MASTER THESIS

Master in Multimedia

Supervisor: Gilberto Bernardes, PhD

Co-Supervisor: Matthew Davies, PhD

July 17, 2019

Soundscape Generation Using Web Audio Archives

Paulo Jorge Fernandes Teixeira

Master in Multimedia

Approved in oral examination by the committee:

Chair: Alexandre Carvalho, PhD

External Examiner: Nuno Fonseca, PhD

Supervisor: Gilberto Bernardes, PhD

July 17, 2019

Abstract

The large and growing archives of audio content on the web have been transforming the sound design practice. In this context, sampling – a fundamental sound design tool – has shifted from mechanical recording to the realms of the copying and cutting on the computer. To effectively browse these large archives and retrieve content became a well-identified problem in Music Information Retrieval, namely through the adoption of audio content-based methodologies. Despite its robustness and effectiveness, current technological solutions rely mostly on (statistical) signal processing methods, whose terminology does attain a level of user-centered explanatory adequacy.

This dissertation advances a novel semantically-oriented strategy for browsing and retrieving audio content, in particular, environmental sounds, from large web audio archives. Ultimately, we aim to streamline the retrieval of user-defined queries to foster a fluid generation of soundscapes. In our work, querying web audio archives is done by affective dimensions that relate to emotional states (e.g., low arousal and low valence) and semantic audio source descriptions (e.g., rain). To this end, we map human annotations of affective dimensions to spectral audio-content descriptions extracted from the signal content. Retrieving new sounds from web archives is then made by specifying a query which combines a point in a 2-dimensional affective plane and semantic tags. A prototype application, MScaper, implements the method in the Ableton Live environment. An evaluation of our research assesses the perceptual soundness of the spectral audio-content descriptors in capturing affective dimensions and the usability of MScaper. The results show that spectral audio features significantly capture affective dimensions and that MScaper has been perceived by expert-users as having excellent usability.

Resumo

Os grandes e crescentes acervos de áudio na web têm transformado a prática do design de som. Neste contexto, sampling – uma ferramenta essencial do design de som – mudou de gravações mecânicas para os domínios da cópia e reprodução no computador. A navegação eficaz nos grandes acervos e a recuperação de conteúdo tornaram-se um problema bem identificado em Music Information Retrieval, nomeadamente através da adoção de metodologias baseadas no conteúdo do áudio. Apesar da sua robustez e eficácia, as soluções tecnológicas atuais assentam principalmente em métodos (estatísticos) de processamento de sinal, cuja terminologia atinge um nível de adequação centrada no utilizador.

Esta dissertação avança uma nova estratégia orientada semanticamente para navegação e recuperação de conteúdo de áudio, em particular, sons ambientais, a partir de grandes acervos de áudio na web. Por fim, pretendemos simplificar a extração de pedidos definidos pelo utilizador para promover uma geração fluida de paisagens sonoras. No nosso trabalho, os pedidos aos acervos de áudio na web são feitos por dimensões afetivas que se relacionam com estados emocionais (exemplo: baixa ativação e baixa valência) e descrições semânticas das fontes de áudio (exemplo: chuva). Para tal, mapeamos as anotações humanas das dimensões afetivas para descrições espectrais de áudio extraídas do conteúdo do sinal. A extração de novos sons dos acervos da web é feita estipulando um pedido que combina um ponto num plano afetivo bidimensional e tags semânticas. A aplicação protótipo, MScaper, implementa o método no ambiente Ableton Live. A avaliação da nossa pesquisa avaliou a confiabilidade perceptual dos descritores espectrais de áudio na captura de dimensões afetivas e a usabilidade da MScaper. Os resultados mostram que as características espectrais do áudio capturam significativamente as dimensões afetivas e que o MScaper foi entendido pelos os utilizadores experientes como tendo excelente usabilidade.

Acknowledgements

To my dissertation supervisor, Gilberto Bernardes, for his huge support, guidance and inspiration over the past few years.

To my co-supervisor, Matthew Davies, for his willingness to help me improving my research and for having integrated me in the Sound and Music Computing group.

To professor Rui Penha for all the incredible moments of sharing knowledge.

A special thank you to all my Friends & AMIDS. It has been an amazing journey just for the simple reason of having you. You are so special. Thank you all!

To my family for believing in me and support me throughout all these years.

Paulo Teixeira

Contents

List of Figures

List of Tables

Abbreviations

API	Application Programming Interface
CSS	Concatenative Sound Synthesis
DAW	Digital Audio Worksation
DR	Dimensionality Reduction
FG	Force-Directed Graphs
M4l	Max for Live
MDS	Multidimensional Scaling
MIR	Music Information Retrieval
MPEG	Moving Pictures Experts Group
MTG	Music Technology Group
PCA	Principal Component Analysis
SC	Star Coordinates
SER	Soundscape Emotion Retrieval
SMC	Sound and Music Computing
SOM	Self-Organizing Maps
SUS	System Usability Scale

Chapter 1

Introduction

1.1 Context

Digital-driven apparatus for recording, manipulating and diffusing sound have significantly redefined the chain of creation, interpretation and reception of music in the XX century [34]. In the 40's, the French composer and engineer Pierre Schaeffer was devoted to the experimentations with sound. The technology of the time allowed Schaeffer to create a new language of musical expression from sound recordings, named concrete music. Hereby, he articulated other studies in the perception and listening domain [59]. Traditionally, the most prominent chain of action in music production to be distributed by digital means relied on recording and synthesizing sound prior to their sequencing in a linear fixed format [4]. At the turn of the XXI century, the massification of music recording and preservation mechanisms, combined with almost unlimited storage capacity offered by modern computing has partially diverted this chain of production and, consequently, music practice [34].

The pervasive availability of private and public (web-based) audio archives points towards a new paradigm of music creation and production, in which the access and retrieval of audio recordings stored on web archives is fundamental [71]. In this context, the audio archive concept has been promoted to a symbol of contemporary culture, thus enforcing its appropriation as a creative prerogative [41]

Music practice and production has been gradually shifting from recording to retrieving audio recordings from archives, due to the large accessibility and ubiquity of this medium [13, 46]. However, new challenges and opportunities were driven from this shift. Audio archives lacked, and to a certain extend still lack, effective retrieval and browsing methods to navigate through their content. Both industry and academia have made efforts in providing algorithms and systems that support the description and retrieval in these vast audio repositories. A prominent example is the MPEG-7 (Motion Picture Experts Group), a standard that provides the description and systematisation of descriptors for multimedia content [30]. With such development, MPEG-7 provides users with the ability to browse, identify and filter audio visual content.

The field of research concerned with the extraction of information from audio signals is commonly known as content-based audio processing within Music Information Retrieval (MIR) field. Its main goal is the extraction of information from audio signal data which can support classification, retrieval or creative transformation tasks [17].

The growth in size and access of digital web audio archives together with the improvements in audio content-based retrieval techniques has been allowing users to more fluidly browse these collections and more accurately retrieve audio recordings. At the intersection of many disciplines such as signal processing, statistics, and machine learning, audio content-based retrieval has been evolving in providing new methods for representing audio signals from a low-level to semantic annotations. These new methods have been supporting a large scope of research across many disciplines such as generative music [49], music recommendation [2] and automatic music transcription [4], to cite just a few.

In the creative realm of music composition, audio content-based methods have promoted a new aesthetic, which emerged from idiosyncratic exploration of audio archives [46]. In the specific domain of our topic, the generation of soundscapes, the use of audio archives has increased in two main directions: 1) supporting a fast access to audio samples, reducing the need for a recording stage [14, 50], and in 2) creating adaptive audio content which blends with some action [6, 49, 65]. Today, both these creative approaches are critical to audio production environments as they streamline the production process as well as respond to current industry challenges in creating adaptive audio and music content for video games, virtual and augmented reality, interactive installations and 360° video [5].

The type of information extracted from audio signals within audio content-based systems is a crucial element in capturing the multidimensional nature of sound and music and thus an important design consideration. The vast majority of content-based retrieval systems extract multidimensional audio data attributes, which is then reduced to a smaller set of information. The advantages of these multidimensional reductions are highly relevant for users. The major advantage of these reductions is the empowerment of the navigation and key insights of the displayed information. Considering that these algorithms exponentially reduce the redundancy of information, which is already a plus, a major advantage is the ability to perceive the most relevant features of each information sample. To this end, the use of dimensionality reduction techniques is typically adopted such as: 1) Principle Component Analysis (PCA) [44], 2) Multidimensional Scaling (MDS) [7], 3) Star Coordinates (SC) [49], 4) Self-Organising Maps (SOM) [47] and 5) Force-Directed Graphs (FG) [54]. Yet, despite revealing the internal structure of the data, most of the resulting representations are dimensionless, i.e., they do not have any attribute assigned to the axis. However, current solutions for navigating audio archives lack a meaningful representation in line with the semantic level of human perception and cognition. There is still a considerable gap between current content-based analysis within audio retrieval methods and human understanding. To promote new strategies for interacting with audio archives, and most notably web audio archives, we present a novel strategy to browse and retrieve audio recordings using a 2-dimensional affective model. Ultimately, we aim to foster a fluid browsing and retrieval method from web audio archives that can assist expert users in the audio multimedia productions.

1.2 Motivation

The main motivation of our work is the growing need to develop a meaningful retrieval method that can assist sound designers in audio production environments, namely in creating soundscapes. We follow an emergent method within the large scope of audio-content based retrieval, which relies on emotion-based soundscape retrieval to support creative tasks. In greater detail, this line of action relies on adjusting the sonic environment according to the emotional states of narrative (e.g. soft city morning to traffic rush hour) could enhance the storytelling and promote the creativity and technical work-flow of expert-users that comprises sound designers, sound editors, foley artists, audio technicians and musicians [72].

Soundscape composition is a pervasive task in sound design production [13]. Typically, soundscape creations are sample-based and the sound materials are often browsed on commercial audio archives or in cloud-based audio archives. The methodology for browsing samples in digital archives is made by scene descriptions, tags or categories; then requiring the audition of each sample in order to find the desired recording, which is an exceptionally tedious task and could break the creative flow while creating soundscapes. Another motivation for this study is the development of a navigable valence-arousal model in order to refine the retrieving process from large audio archives. Many experiments were done in MIR towards an effective browsing but only a few have been adapted into the current audio production industry.

Most of the recent developments in soundscape field of research focused on extending the taxonomical organisation of sounds, data retrieving techniques and the automatic generation of sonic content. The present study followed the related projects and advances on the field which allowed us to research about the online audio archives as potential tool for artistic explorations and the development of a navigation model for better retrieving and generation of soundscapes using online audio archives.

1.3 Project

The proliferation of audio through web archives has changed the way we interact with audio data. The emergence of archive's aesthetics required a meticulous control of how we: 1) retrieve, 2) dynamically navigate and 3) represent data; expanding the potential of audio uses for sonic creation. Most of the available soundscapes systems do not satisfy all the three aforementioned requisites. The scope of this dissertation is the exploration of online audio archives for creative sonic experiences and consequent development of a system for efficient retrieval of soundscape audio recordings from Freesound - MScaper. This application aims to leverage the potential of a cloud-based repository such as Freesound to assist expert-users when generating soundscapes for games, TV, films and installations. For this purpose, users can query Freesound to retrieve sets of sound objects that compose the soundscape according to Schafer's taxonomy [62]. Besides, MScaper supports the navigation through each collection of sound objects using a valence and arousal model, granting an adaptive soundscape generation according to emotional states of the user. The project aims to shed light on how users can perform, compose and interact with online audio archives through the development of a framework for bi-directional communication with Freesound, Fs.Library.

We can formulate our problem considering the current limitations:

- Lack of non-linear soundscape systems using web audio archives;
- Poor levels of representation and navigation for retrieving audio recordings.

Problem:

- How to leverage the potential of web-based audio archives for soundscape creation?
- How to design a meaningful navigable space for retrieving audio recordings?

Hypothesis:

- Hypothesis is that the valence and arousal models from low-level audio description can foster an intuitive navigation model for retrieval and assist the dynamic creation of the soundscapes.

Objectives:

1. To understand how Freesound can streamline and assist the dynamic soundscape generation;
2. To study the validity of a valence-arousal model from low-level audio description for retrieving audio recordings from Freesound and adaption of the soundscape.

1.4 Dissertation Structure

This dissertation is structured in five chapters. The chapter 1 presents the introduction, the context and motivation and defines the objectives of the research. Chapter 2 encompasses the literature review. It starts by presenting two authors and describing their theoretical listening modes, followed by the review of state-of-the-art audio description schemes and musical spaces. The chapter 2 concludes with presenting and comparing some of the most relevant generative audio systems in the panorama. The chapter 3 aims to detail a mapping between perceived affective dimensions and audio spectral descriptions. In chapter 4, is presented the library to interact with Freesound and the developed Max for Live (M4L) application device. It starts by breaking down the set of Max MSP modules and the development of the valence and arousal model. In chapter 5, the evaluation of the research is presented. It starts by presenting the audio features analysis of the audio dataset used for the Valance-Arousal model and then the tests made using the MScaper application, a perceptual test and system usability test. Finally, chapter 6 presents the conclusions and contributions of this work followed by a commentary on future directions.

1.5 Publication

This research led to the following paper:

- Teixeira, P., Davies, M., Bernardes, G. (2019). The Online Musical Database in / as Performance. Hidden Archives, Hidden Practices: Debates About Music-Making. Aveiro, 2019.

Chapter 2

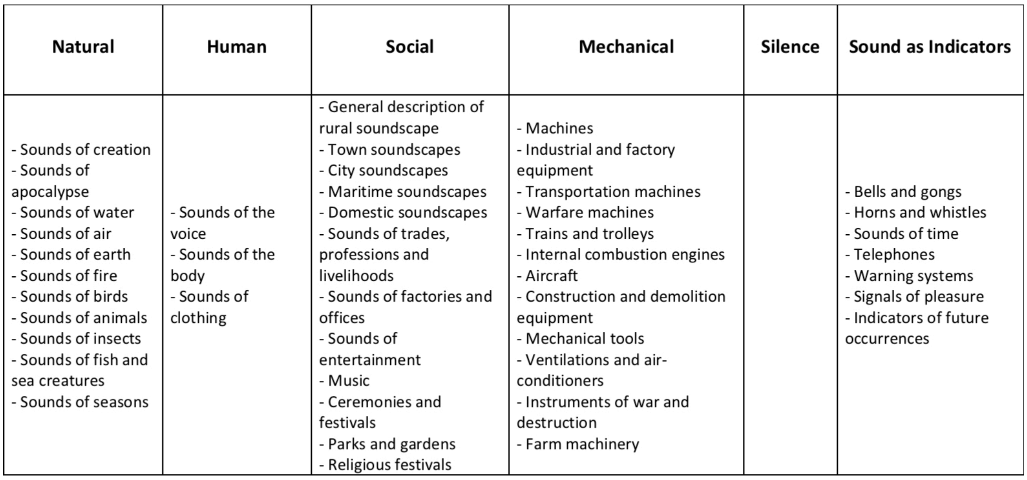
From Computer Audition to the Procedural Generation of Soundscapes: A State-of-the-Art Review

2.1 Soundscapes

The contributions of concrete music aroused interest to the World Soundscape Project, a research group established by Murray Schafer at the Simon Fraser University. Concerned about acoustic ecology of spaces and their preservation, the group considered the rapidly changing sonic environment of spaces and noise pollution a subject of research, as presented in “The New Soundscape” [60] and “The Book of Noise” [61].

In 1970s, Murray Schafer defined soundscape as: “an acoustic environment or an environment created by sound. The sonic environment. Technically, any portion of the sonic environment regarded as a field for study. The term may refer to actual environments, or to abstract constructions such as musical compositions and tape montages, particularly when considered as an environment.” ([62], p. 274). The Soundscape movement started in late 1970s when Schafer presented the book “The Soundscape: Our Sonic Environment the Tuning of the World” ([62]). In this book, Schafer assembles the previous research done within the World Soundscape Project [60] about the acoustic overload of the 60’s and presents the soundscape concept.

The classification of soundscapes and their sound objects established in Schafer’s method is crucial to discover similarities, contrasts or patterns. Schafer distinguishes three main structures of a soundscape: keynote sounds, signals and soundmarks. The keynote sounds are analogous to the musical notion of key within tonal music, the reference point or anchor of a composition. The author compares the concept of keynote sounds in soundscapes to the visual perception of figure and ground “as figure and ground are contrasted in visual perception.” ([62], p. 275). Signals are sounds in the foreground that draw the listener’s attention. Theoretically, any sound can become a signal sound if the listener decides to pay attention to it. Schafer ([62]) further defines the signal sound as those sounds that force attention from the listener, such as bells, horns and whistles. Soundmark sounds are those particularly noticeable for a distinct community, deriving from the term landmark. The referential classification of sound sources described by R. Murray Schafer can be seen in the Figure 2.1 bellow.



Natural	Human	Social	Mechanical	Silence	Sound as Indicators
- Sounds of creation - Sounds of apocalypse - Sounds of water - Sounds of air - Sounds of earth - Sounds of fire - Sounds of birds - Sounds of animals - Sounds of insects - Sounds of fish and sea creatures - Sounds of seasons	- Sounds of the voice - Sounds of the body - Sounds of clothing	- General description of rural soundscape - Town soundscapes - City soundscapes - Maritime soundscapes - Domestic soundscapes - Sounds of trades, professions and livelihoods - Sounds of factories and offices - Sounds of entertainment - Music - Ceremonies and festivals - Parks and gardens - Religious festivals	- Machines - Industrial and factory equipment - Transportation machines - Warfare machines - Trains and trolleys - Internal combustion engines - Aircraft - Construction and demolition equipment - Mechanical tools - Ventilations and air-conditioners - Instruments of war and destruction - Farm machinery		- Bells and gongs - Horns and whistles - Sounds of time - Telephones - Warning systems - Signals of pleasure - Indicators of future occurrences

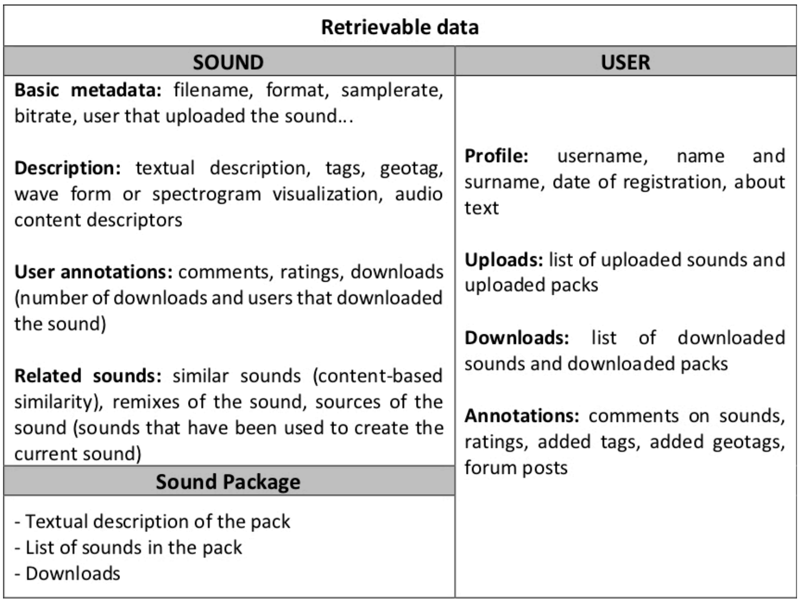
Figure 2.1: Sound classification proposed by Murray Schafer [62].

2.2 Freesound

Freesound is an online audio archive for sharing recorded audio clips under the Creative Commons license. It started in 2005 and it is being further developed by the Music Technology Group (MTG) of the Universitat Pompeu Fabra. The increasing number of audio content available on Freesound demanded the development of a technological structure allowing rapid data retrieval and filtering. Considering the archives as a mean for artistic practices, Freesound and the possibility of retrieving data through an Application Programming Interface (API) has been used for artistic research in many different areas, such as: soundscape generation ([24], [70], and [19]), sound browsers ([14] and Timbral Explorer), creative interfaces and composition ([45], [77] and [71]).

2.2.1 Database Interaction

The Freesound API allow users to interact with content on platform by browsing, searching, filtering and retrieving information about sound samples, sound packages and Freesound users [1]. It allows deeper comparisons between content by searching similar sounds to a given target based on content analysis and retrieving automatically extracted features from audio files. Furthermore, the API allows to browse by combining content analysis and other metadata (geolocation, tags) as depicted in Figure 2.2.



Retrievable data	
SOUND	**USER**
Basic metadata: filename, format, samplerate, bitrate, user that uploaded the sound... **Description:** textual description, tags, geotag, wave form or spectrogram visualization, audio content descriptors **User annotations:** comments, ratings, downloads (number of downloads and users that downloaded the sound) **Related sounds:** similar sounds (content-based similarity), remixes of the sound, sources of the sound (sounds that have been used to create the current sound)	**Profile:** username, name and surname, date of registration, about text **Uploads:** list of uploaded sounds and uploaded packs **Downloads:** list of downloaded sounds and downloaded packs **Annotations:** comments on sounds, ratings, added tags, added geotags, forum posts
Sound Package	
- Textual description of the pack - List of sounds in the pack - Downloads	

Figure 2.2: Retrievable content from Freesound [1]

Freesound archive is one of the biggest sound archives in which content is prepared to be accessed through an API, contributing to the development of third-party applications that explore the audio archives as a creative tool. These applications communicate with Freesound using client libraries for Python, JavaScript, Perl, Supercollider, Objective-c, Qt framework and Scala. The Freesound Explorer [25] is a visual interface to explore Freesound content in a two- dimensional space and aims musical composition while discovering the Freesound. The system benefits from some services of the API like query search, content-based similarity and timbral properties analysis.

2.2.2 Essentia

The possibility to analyse and retrieve data from audio content using Essentia is one of the main Freesound API attributes that enhances the ability to use Freesound for research or musical practices. Essentia is a “cross-platform open-source library for audio analysis and audio-based music information research and development” [9], focused on robustness and performance on large audio archives. It contains a set of algorithms which implement audio input/output functionality, standard digital signal processing blocks, filters, algorithms for statistical characterisation of data and a large set of spectral, temporal, tonal and high-level music descriptors.

The Essentia library is broadly used in new industries that leverage the use of technologies developed in MIR, such as: music recommendation (BMAT and Stromatolite) and automatic playlist generation (Yamaha).

2.3 Computational Audio Description

The ever-increasing digital audio on big cloud-based audio archives required the development of technologies capable of extracting information from audio content, providing new ways for browsing and interacting with large collaborative unstructured archives. Most content-based audio processing research focuses on the recognition of sound sources [29], music classification [40], and music recommendation [12].

Early developments on content-based processing started with the development of MPEG-7 framework, a standard for multimedia content description. This standard defines low-level description techniques as well as high-level tools for multimedia management [30], allowing users to search, identify and browse audiovisual content. The CUIDADO (Content-based Unified Interfaces and Descriptors for Audio/music Databases available Online) project explored MPEG-7 content description Interface to develop content-based technologies to create the next generation of audio content management systems [74].

Throughout the last two decades, MIR research field has contributed with novel ways to understand and describe sounds [17]. The primary goals of MIR are the extraction and interpretation of features from music, music-indexing using these features and search and retrieval schemes, as stated by Downie (as cited in [17]).

2.3.1 Audio Descriptors

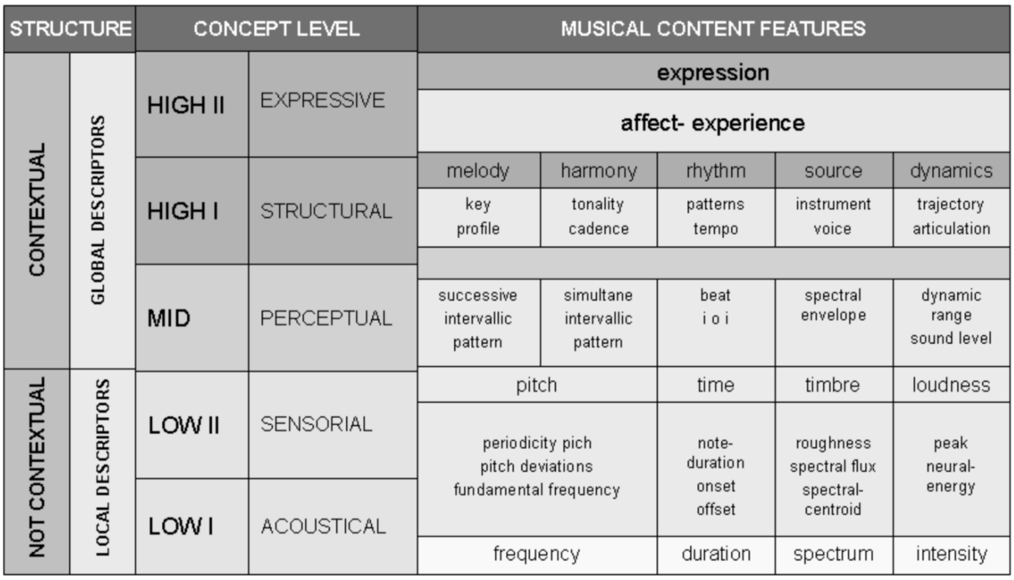
Audio descriptors are technologies used for representing audio by measuring certain properties of audio signal content and folding them to a range of values. The selection of appropriate audio data representation when managing content-based retrieval systems is fundamental since these are the mean used to interact and browse on large audio archives.

According to Schwarz [64], audio descriptors can be organised into three different classes: 1) categorical, descriptors that belong to a class; 2) static, descriptors that measures a value for the whole signal, for example the attack duration of a sound; 3) dynamic, descriptors computed for each sample or other time frame, for instance the spectral centroid of a signal that can vary along the time. Schwarz [64] complements referring another method to organise audio descriptors according to their level of abstraction. These levels are the following:

- **Low-level**: retrieved directly from various signal processing techniques, either on the time domain such as amplitude and zero-crossing rate; or on the frequency domain like spectral centroid and spectral flatness [64];
- **Mid-level:** require some level of interpretation from the data. This group of descriptors infers information directly from the audio data or from the results of a prior analysis. Common mid-level descriptions are harmony, key and rhythm;
- **High-Level:** user-centered that operate at a semantic level, providing information on how humans understand and interpret music like genre, mood or style. The distance between mid and high-level description is often referred as semantic gap since the labels from high-level descriptors

cannot be promptly obtained from lower level features.

Common audio retrieval systems deal with the filtering of several low-level signal description values. Although, when it comes to explore large archives for creative purposes there is a need to search content according to meaningful sonic dimensions to the users. Lesaffre et al. [3] developed a user-centered taxonomy and framework for audio description. The author considered the importance of providing a common language of descriptors to the users and developers of audio information retrieval systems. The structure distinguishes two types of descriptors, namely local and global descriptors. The proposed taxonomy (See Figure 2.3) comprises the aforementioned levels of description in a multi-leveled taxonomy and categorise them according to the musical content features, low (physical and sensorial), mid (perceptual) and high description levels (formal and expressive). The structure of the taxonomy also includes six categories: melody, harmony, rhythm, timbre, dynamics and expression.



STRUCTURE		CONCEPT LEVEL		MUSICAL CONTENT FEATURES				
CONTEXTUAL	GLOBAL DESCRIPTORS	HIGH II	EXPRESSIVE	expression				
				affect- experience				
		HIGH I	STRUCTURAL	melody	harmony	rhythm	source	dynamics
				key profile	tonality cadence	patterns tempo	instrument voice	trajectory articulation
		MID	PERCEPTUAL	successive intervallic pattern	simultane intervallic pattern	beat i o i	spectral envelope	dynamic range sound level
				pitch		time	timbre	loudness
NOT CONTEXTUAL	LOCAL DESCRIPTORS	LOW II	SENSORIAL	periodicity pich pitch deviations fundamental frequency		note- duration onset offset	roughness spectral flux spectral- centroid	peak neural- energy
		LOW I	ACOUSTICAL	frequency		duration	spectrum	intensity

Figure 2.3: User-Oriented Approach to Music Information Retrieval [3]

2.3.2 Perceptual Audio Description

The browsing process in content-based audio retrieval systems is made by querying large archives using audio features as parameters, proving to be a very efficient method as seen in CUIDADO project [74]. However, there have been major efforts in the MIR community to foster the semantic degree and consequently the usability of the taxonomies for audio retrieval schemes. Currently, audio descriptors are often based on statistical methods and apply a rather technical terminology, which is hard to grasp for those without a scientific or engineering background. Perceptual audio descriptions aim to capture how listeners perceive and understand audio, as well as adopt a rather 'musical'-oriented terminology of semantic level, such as loud and soft dynamics, instrumental timbres (e.g., bright or dark color) and pitch. These descriptions are more user-friendly given the semantic level of the terminology. As an example of this terminological gap we can establish a link between a similar phenomena describing sound as ’bright’, instead of a sound with a high spectral centroid. However, from a descriptive computational approach, and perceptually for that matter, they are equivalent [63].

In order to describe perceptual attributes of musical timbre, users usually tend to rely on semantic descriptions such as bright, resonant or clean, which are features correlated to timbre [73]. A contribution towards common terminology of signal description is described in Grill’s framework for perceptual audio description [27]. Grill studied how people could use a common terminology to describe sounds leading off by finding what were the best notions to describe sound attributes. The experiment resulted in descriptors that portray spectral qualities such as: high-low, tonal–noisy; and structural-temporal, ordered–chaotic, smooth–coarse, homogeneous–heterogeneous aspects of sound. Other contribution to common terminology of signal description is described in Analysis Toolkit for Sound-Based Composition [8]. This musicological description approach let users identify sound objects from audio and to interface sound objects descriptions adapted to music lexicon. Based on Schaeffer’s [59] perceptual criteria and Peeters [52] information redundancy across audio descriptors, the description scheme adopts a reduced number of descriptors, selected based on perceptual criteria for sound description.

Regarding semantic timbral descriptions and their effectiveness on searching process through large sound archives, there have been successful attempts in MIR. Pierce [50] considered what were the meaningful semantic timbral attributes to users gathering the frequency of textual search of timbral attributes by Freesound users. This study observed that brightness, depth, hardness and roughness are the most semantic timbral qualities searched in Freesound. The outcomes from this study can be leveraged to achieve more accurate audio retrieval in big audio archives, as seen in Perceptual Sound Browser [14]. Recently, Pierce [53] launched a new version of Timbral Explorer, previously exhibited in the context of Abbey Road Hackathon (2018). The project consists in a web-based app that let users query sound objects to Freesound (first version used fixed classes) and then search results are distributed across the screen in a 2D plane according to two semantic timbral descriptions assigned to each axis, providing an instinctive visualisation of sonic content.

2.3.3 Description Spaces

The purpose of description spaces is to reveal insights on the data content through graphical representations. The most common way to represent archives content in audio processing applications is to include a large number of audio features in a two-dimensional space providing an intuitive navigation through an archive of audio recordings [49]. Concatenative sound synthesis (CSS) has seen many successful applications using description spaces, being one of the earliest well-known systems CataRT for Max/MSP [65]. The description space of CataRT is reduced to a two-dimensional projection according to two selectable descriptors assigned to each axis and uses the mouse to explore an archive of audio recordings. The need for an effective representation of audio content using more than two features for more consistent exploration of sounds gave rise to dimensionality reduction (DR) techniques in audio applications. The DR algorithms decrease the number of dimensions while retaining most of the information provided by the original vector. The idea is that the distances between two points in the high-dimensional vectors are preserved in the lower-dimensional one, leveraging the features analysis. Description spaces in content-based audio processing applications foster an intuitive understanding of the audio content and streamlines the sonic creation process.

Self-Organizing Maps (SOM) reduction technique is a non-supervised learning algorithm as well as a powerful visualisation tool [15]. In Islands of music [47], similarity between pieces of music is estimated considering timbre, rhythm patterns and metadata information. Then, pieces are organised in a two-dimensional space where the similar pieces are grouped together. A visualisation using a representation of geographic maps provides an intuitive interface where "islands" are meant to be genres or styles of music.

Bernardes [49] presents EarGram, a Pure Data application for real-time CSS. The system incorporates four generative music strategies that rearrange and explore an archive of audio recordings. EarGram implements the Star Coordinates DR technique [32] that maps a high-dimensional data linearly to two dimensions by summing the vectors resulting from the point coordinates arranged on a circle on a two-dimensional plane with equal (initially) angles between the axes with an origin at the center of the circle. The visualisations, similar to CataRT, may assist the musical paths during performance [49]). Another DR technique, and probably the most well-known algorithm, is the Principal Component Analysis (PCA), which expresses each one of the dimensions in the latent space as a linear combination of all the dimensions in the higher space. Nuanáin [44] describes an interactive CSS instrument that generates and visualises rhythmic patterns from existing audio in real-time. The DR is applied on each feature vector of the units in a loaded selection of audio files. Two PCA components are retained and scaled to the visible area of the interface to serve as coordinates for placing a circular representation of the sound in two-dimensional space [44]. These visual representations provide information about audio content allowing coherent explorations in archives.

Multidimensional Scaling (MDS) technique focuses to represent similarities or dissimilarities in the form of a geometrical model by organising the elements in a coordinated space [21] being a good method for visualising similarities between sounds. This algorithm was employed in MixMash [7] to create a hierarchical harmonic mixing method for assisting users in the process of music creation. MixMash [54] is an interactive tool to aid in music mashups through associations between musical content analysis and information visualisation. This system leverages the harmonic mixing method for music rearrange from the aforementioned EarGram [49] and use force-directed graphs to represent the good and less-good harmonic compatibility. The force-directed graph visualisations are a good technique to handle the representation of relational structures in data using 1) nodes, to represent a class and 2) edges to represent the relationship between nodes. In Mix Mash's interface, nodes represent audio tracks and musical keys. The distance of tracks and keys define the weight for each edge and therefore the force of attraction between nodes. The less distance of nodes, higher force of attraction and consequently, closer nodes. These forces are applied to the nodes enabling easier interpretation of the graph and interaction when searching for harmonically compatible tracks.

Description spaces navigation is the most common way to interact large sound archives with high degree of content comprehension, being the way data is clustered and the how we interact with it, fundamental for coherent sonic outputs. For automatic generation of soundscapes, these spaces are responsible to foster an intuitive navigation through large archives and the ability to dynamically adapt the soundscape in real-time according to the composer's ideas. The projects presented above employ multidimensional reduction algorithms in order to reduce a high-dimensional vector into a lower one, usually two or three dimensions. From literature review, the use of these algorithms is generally useful for data visualisation exploration. Bernardes [49] referred that the comprehension of the axes in SC takes places over the PCA and MDS algorithms, on the other hand, a clear drawback "is the need to explore the representation by weighing the variables and assigning different angles to the axes" (p. 87). SOM is well adapted for handling huge amounts of data, commonly used in economy and industrial sectors [15], which makes it an interesting choice to explore large audio archives. However, a common weakness of these visualisations is the need to learn through the exploration of the system considering the lack of a direct link to high-level perception of sound. Valence-Arousal models are a powerful method for navigation on large audio archives as well as to promptly link users to audio emotional descriptions filling some of the gaps of the aforementioned techniques. Even though there is an absence of automatic soundscape systems using this model, the valence-arousal has showed very effective in music mood recognition [66] and music recommendation [2]).

2.4 Generative Soundscape Systems

The automatic generation of sonic content is one of the most typical research topics within the Sound and Music Computing (SMC) field. Thus, this part of my dissertation aims to review only the most relevant projects to my research and their contributions to the state of the art followed by a comparison of each application's contributions. This section presents two broad categories of generative systems. The first part focuses on the soundscape generative systems that use personal audio collections. The second part describes systems that retrieve audio recordings from web audio archives and use them for soundscapes creation.

The selection of the projects is based on the following criteria:

- Employ an audio archive or allow the management of multiple audio files;
- Dynamic generation of soundscapes through navigation or sound manipulation;
- Most well-known generative systems regarding the techniques used and public target.

TAPESTREA

Musical Tapestry: Re-composing Natural Sounds [42]

TAPESTREA stands for Techniques And Paradigms for Expressive Synthesis, Transformation and Rendering of Environmental Audio. TAPESTREA is a technique and system that rearranges recorded sounds by separating them into unique components and merging these components into musical tapestries. The system aims to assist users while creating soundscapes through analysis, transformation and re-synthesis of natural sounds. For this purpose, the application employs subtractive synthesis techniques to isolate and extract sounds selected by the user and to extract residual background which is further parameterized by wavelet tree analysis. All these extracted elements are further sequenced into a graphical interface that allows the user to manipulate the generated soundscape through the application environment or by scripting in Chuck programming.

EarGram

Application for Interactive Exploration of Concatenative Sound Synthesis [49]

EarGram is a CSS application built in Pure Data for interactive exploration and creation of sonic content. The system gathers the potential of content-based analysis of an audio archive and different methods of visualization of audio content along with an user-friendly terminology to dynamic creation of sound. The application encompasses four generative strategies for interactive music contexts: SpaceMap, InfiniteMode, SuffMeter and SoundscapeMode. This last recombination method aims to synthesize soundscapes in real-time by spreading sound segments according to perceptual attributes in a two-dimensional navigable space. The soundscape mode also provides the manipulation of two set of controls: 1) the density of events (dense and sparse), and 2) roughness of the events (smooth and sharp); providing dynamic generation of soundscapes.

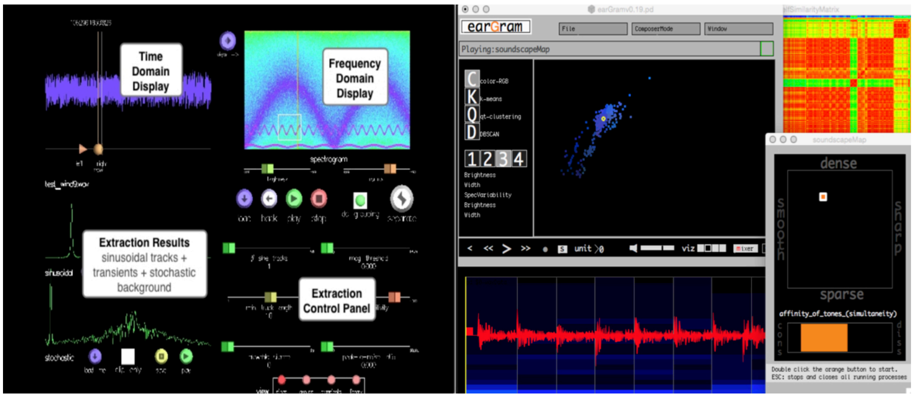


Figure 2.4: Screenshots of the TAPESTREA (on the left) and EarGram (on the right) generative systems.

TAPESTREA and EarGram

I consider both, TAPESTREA and EarGram, very meaningful applications to my research and to the general panorama of SMC state of the art. These systems differ concerning their target audience, user experience and its methods for sonic generation, yet both allow dynamic generation of soundscapes. The main difference between these applications is their target audience and their efficacy in different contexts. TAPESTREA uses very useful sound creation methods for those who are expert-users, applying technical terminologies used in audio production environments, while EarGram employs a more musicological vocabulary being more user-friendly. Another main difference is the way users interact with the system in order to generate soundscapes. In one hand, TAPESTREA has a very familiar interface to sound designers, allowing to sequence a timeline with individual sound elements. A clear advantage of TAPESTREA over EarGram is the possibility to extract individual sound objects from a recording and use them to create another scenario. On the other hand, EarGram's interface provides an intuitive understanding of perceptual audio qualities spread in two-dimensional screen, allowing great interaction with an audio archive. All in all, I consider that the major contributions from both applications is the generative strategies. Specific contributions from both is EarGram's navigable space in audio archives and the sound modelling techniques from TAPESTREA.

Coming Toghether

Negotiated Content [18]

A generative soundscape system where four autonomous artificial agents choose audio recordings from a large analysed archive of soundscape recordings from Freesound, based upon on their spectral content and metadata. Regarding the spectral selection, agents interact during the performance with others by analysing the spectral content of each individual agent, avoiding spectral overlapping. The metadata selection is provided by the agent's listening perception of the sonic environment, consequently related metadata is generated. As part of Metacreation Labs, this system aims compose soundscape compositions that would be considered creative if generated by a human. Negotiated content is a soundscape composition using this system. During the composition, the above-mentioned selection methods vary: during the first of four sections, agents select material based entirely upon metadata tags; during the final two sections, agents select material based entirely upon spectral regions; during the second section, both methods are used. Finally, each agent adds granulated instrumental tones at the resonant frequencies.

API.Cultor

Sound recycling from public databases [45]

API.cultor is a musical system oriented for live performances and experimental music composition. This musical interface aims to use analysis techniques from MIR combined with web audio archives, dynamic user interfaces, physical controllers and real-time synthesis, keeping in mind compositional concepts and focusing on artistic performances. This collaborative system flows by retrieving sounds from Fresound and RedPanal archives, using REST API, and then trigger these sounds by computer and cellphones through network or using the Pi-based API.Cultor instrument. The real-time synthesis is also controlled by the users during the performance. The system allows to perform in two modes: deterministic and stochastic. The main difference between them is that in stochastic mode, every time a sample is retrieved, there are many candidates to be played, being the system responsible for randomly choose one.

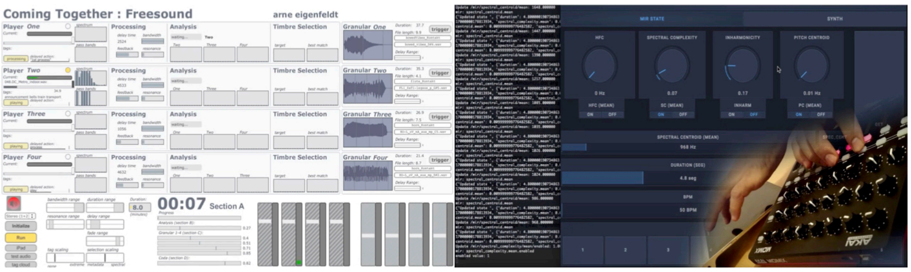


Figure 2.5: Screenshots of Coming Together : Freesound (on the left) and API.Cultor (on the right)

Coming Together and API.Cultor

The aforementioned projects are both good references to understand the possibilities of working with an audio archive for the generation of sonic content, namely soundscapes. In both, the development of new ways of live collaboration with musical goals and the exploitation of available online resources seems to be effective to achieve new ways of thinking music. To note, the API.Cultor's system subtheme is "Sound Recycling from Public Databases" which appeals to the aesthetics of using audio archives as mean to artistic purposes. Both projects can be considered real-time generative systems for soundscape composition, meant to be used in live performances. However, they differ in their usability, interaction and the role of the performer. In Coming Together, the composer is aided by four autonomous agents for sonic generation during performance, being the composer responsible to initially feed the system and the agents play/compose along. The same level of indeterminism is seen in API.Cultor since the system can perform instructions to retrieve and play sound from devices connected to the network.

2.5 Summary

In this Chapter, I provided a general overview of the most relevant state of the art studies in soundscape retrieval and generative methods, namely those relying on web audio archives. There are many articles aiming to describe audio regarding how listeners perceive sound and techniques to improve how can users visualise and navigate in web audio archives for generation of soundscapes.

Despite the satisfactory results, there is a huge gap between the number of projects dedicated to search, retrieve and describe audio recordings from web audio archives to generative systems using those web services. Furthermore, there is a lack of systems offering an intuitive navigation model for retrieval and dynamic adaption of the soundscape in real-time.

Chapter 3

Mapping Affective Soundscape Dimensions to Low-level Audio Descriptions

This Chapter details the mapping between perceived affective dimensions and audio spectral descriptions in soundscapes recordings. Its ultimate goal is to provide an audio descriptors fingerprint that captures affective dimensions in Russel's circumplex model [55]. To this end, in Section 3.1, we start by detailing Russel's circumplex model of emotions. In Section 3.2, we present the dataset Emo-Soundscapes, which provides crowdsourced affective annotations for soundscapes in the Russel circumplex model of affect. We then identify the most prevailing affective spaces in the dataset through the adoption of data mining and visualisation methods. In Section 3.3, we present a strategy to quantize the affective space into a discrete equal-spaced square matrix. Additionally, we define the concept of audio fingerprint with vectors of descriptions for each quantized space in the plane. Finally, in Section 3.3.2, we provide some statistical analysis of the audio content-based descriptors which best express the affective distribution of soundscapes in the space.

3.1 Russel's Circumplex Model of Affective States

A long-identified problem in the field of affective psychology is the assessment of emotions [58]. The problem emerges from our perception of emotions as ambiguous and intercorrelated, rather than isolated and discrete. Similar to the spectrum of colours, Russel [57] claims that the lack of discrete borders inept us to clearly distinguish one emotion from another.

In greater detail, affective psychology has showed that people describe an affective experience by referring to more than one emotion [75]. Furthermore, research has also indicated that affective experiences are intercorrelated both within and between people [56]. These intercorrelations between emotions are notably addressed in dimensional affective models [68, 55]. These models express affective experiences as a continuum of correlated ambiguous states[56].

Research on the intercorrelations among affective experiences has repetitively employed two-dimensional representation models. These two-dimensional models use dimensionality reduction techniques, such as multidimensional scaling and factor analysis [37], to reduce the multidimensional nature of the phenomena and retain the most relevant information. Different authors have conceptualised these dimensions as: positive and negative affect [76], tension and energy [68], approach and withdrawal [36] and valence and arousal [55]. Regardless of the contrasting names of these dimensions, they point towards a similar twofold continuous structure. In other words, independently of the adopted dimensions, these circumplex models of affect assume that all affective states result from two independent neurophysiological dimensions.

A prominent circumplex model of affect is by Russell [55]. All affective states can be understood as a linear combination of two independent dimensions. In detail, points in the space express affective states as varying degrees of valence and arousal (see Figure 3.1). Valence, in the *x*-axis, expresses a continuum of pleasantness that an event evokes. It ranges from negative to positive, or from displeasure to pleasure. Arousal or alertness, in the *y*-axis, is the perceived activity of an affective event and ranges from calm to excited.

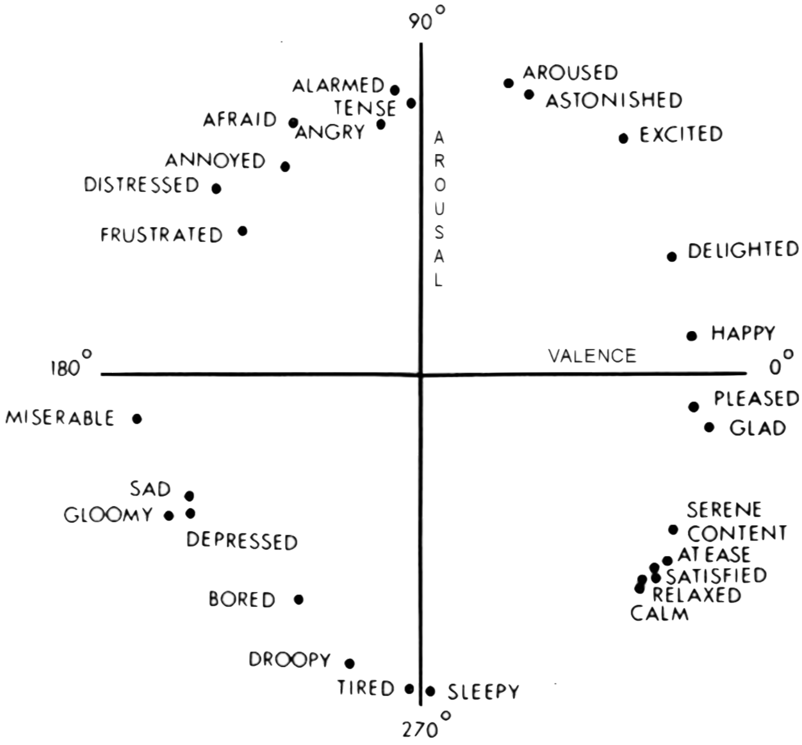


Figure 3.1: Emotional plane - Russell's Circumplex Model [55]

Specific emotions are then defined by these these two dimensions. Happiness, for example, is conceptualised as an emotional state that is the combination of positive valence together with moderate arousal. Misery, on the other hand, results from negative valence and moderate arousal.

3.2 Emo-Soundscapes: Affective Annotations of Soundscape Sounds

There have been major efforts in the last decade from the MIR community to present robust models that automatically identify affective states from audio content [26]. Particular emphasis has been given to speech [33, 28] and musical signals [16, 48]. Recently, the scope of studies on emotion retrieval from audio signals has been extended to soundscape sounds. In this context, an important contribution as been the Emo-Soundscapes dataset [22], which comprises a 1213 6-second Creative Commons licensed soundscape sounds with annotations of perceived levels of valence and arousal and multiple low-level audio descriptions.

In the Emo-Soundscapes dataset, 600 soundscape sounds were retrieved from `Freesound.org`[1] and the remaining 613 soundscape sounds result from different mixes of the retrieved set of soundscape sounds. The selection process from Freesound was curated by three soundscape composers following a balanced number of 100 audio recordings for each of the six proposed soundscape categories by Schafer [62], previously detailed in 2.1. The referential classification categories are the following:

- Natural;
- Human;
- Social;
- Mechanical;
- Silence and Quiet;
- Sound as Indicators.

The *Silence and Quiet* category lacks a link to a context or an identifiable source. For example, in Schafer's taxonomy, a quiet park is categorised under quiet and silence. However, it could also be classified as natural sound.

Each retrieved audio recording from Freesound was trimmed to a 6-second duration using an automatic segmentation strategy. The segmentation strategy aimed to retain the most prominent event in the soundscape [69]. This algorithm targets a background and foreground separation and segmentation, from which is then selected a continuous 6-second excerpt that best relates to the semantic tags associated with the soundscape.

To study the impact of the perceived arousal and valance on soundscape mixes, the remaining 613 sounds in the dataset resulted from manually combining two or three soundscapes from the first subset. To this end, soundscapes within and across Schafer's categories are mixed. Furthermore, the impact of loudness, or balance in mixing of the soundscape files was considered [22].

[1] Freesound.org is an online sound database where people share recorded sound recordings under Creative Commons licenses.

Annotations of perceived affect have been collected for each of the 1213 soundscapes using the CrowdFlower[2] online platform. In total, 1182 *trusted annotators* [22] from 74 countries have provided perceived degrees of valence and arousal in Russel's 2-dimensional circumplex model [55].

Every participant had to rank a minimum of 5 pairs of audio clips before completing the collection of affective data and being payed. To avoid uneven distribution of soundscape annotations, a parallel ranking method in which three corpora of audio recordings are ranked in parallel was adopted.

Emo-Soundscapes Metadata	**Example**
File Name	$\mathrm{r}_0 human101979_1 313063 - hq.mp3$
Search Term	cry
Duration	6 seconds
Class	Background
Sound ID	101979
URL	http://www.freesound.org/people/Raffa%20Jaffa/sounds/101979/
Tags	cry, field, istanbul, man, prayer, recording,temple

Table 3.1: Example: Metadata Available for Every Emo-Soundscapes Audio Recordings

Besides the affect (valence and arousal) annotations, the dataset includes audio content analysis and the following metadata per soundscape: 1) file name, 2) search term (within Schafer's taxonomy), 3) duration of the original file, 4) class (background sound, foreground sound, background with foreground sound), 5) sound identifier (ID) in Freesound, 6) URL and 7) their Freesound-associated tags. The Table 3.1 presents, as an example, the metadata available for the first audio recording from Emo-Soundscapes.

3.2.1 Perceived Affective Annotations

To unpack information driven from the collected crowdsourced affective data from the Emo-Soundscapes dataset, we explore in Figure 3.4 the visualisation of the dataset in the Russel's circumplex model. Each of the 1213 points in Figure 3.4 corresponds to a soundscape sound in the Emo-Soundscapes dataset. Their location in the 2-dimensional space results from applying the Klaus Krippendorff algorithm [35].

The distribution of the annotations in the 2-dimensional valence-arousal space is uneven. There is a high density of points in the diagonal that links quadrants one and four (low valence/high arousal and high valence/ low arousal). The dense diagonal distribution of the Emo-Soundscapes audio recordings seems to suggest that the affective perception of soundscapes occurs mostly between the continuum that ranges from distressed to calm. Contrarily, the remaining quadrants three and four present a rather small number of points. [31, 39]

Research studies of music emotion retrieval have been using automatic models to classify emotions from audio signals[31, 39]. These models, regressors, has a benefit of the continuous

[2] https://www.crowdflower.com/, accessed on June 21st, 2019

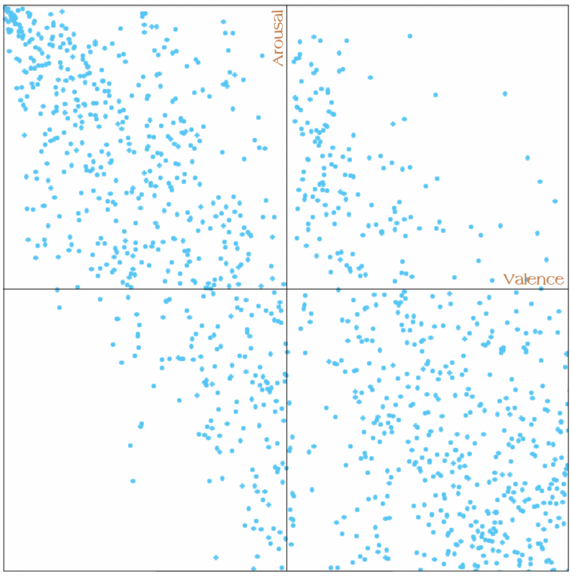


Figure 3.2: Visualisation of the affective valence-arousal crowdsourced annotations of Emo-Soundscapes dataset.

approach is in the application domain: it makes it possible to generate a playlist that transitions smoothly from one emotion to another by following a path in the AV plane. In contrast to SER, the most prevalent space in the affective plane, given its density, is the diagonal that ranges from sad to excited.

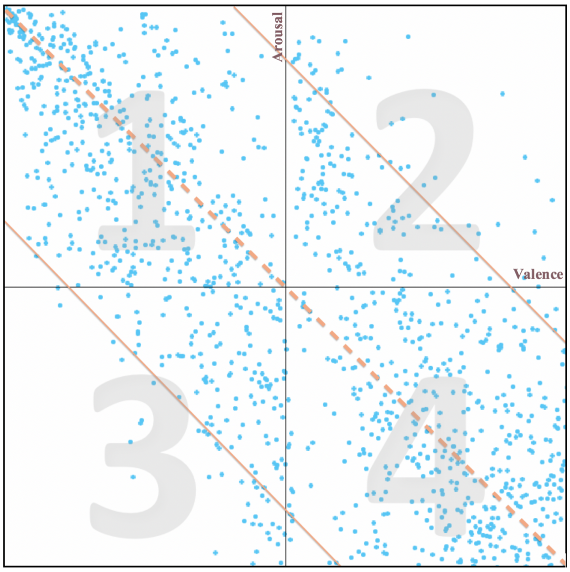


Figure 3.3: The prominent space marked with a dashed diagonal. Every quadrant of the plane is represented by a number

Despite being less prominent than the aforementioned diagonal, the vertical and horizontal axis that divide the plane into four quadrants are also well represented in the annotation space. On the *x*-axis, the existing soundscape recordings are expressed between the continuum that ranges from unpleasant to pleasant. In the y-axis, soundscapes exist between the continuum that ranges from sleepy to tense.

3.2.2 Low-level Audio Descriptors

In addition to the affective annotation, the Emo-Soundscapes includes low-level content-based analysis for each soundscape in the dataset. The MIRTollbox [38] and YAAFE [20] libraries were adopted to extract the audio descriptor data.

Temporal Descriptors	**Spectral Shape Descriptors**	**Energy Descriptors**	**Harmonic Descriptors**
Zero crossing rate	Spectral Flux	Low Energy	Inharmonicity
	Spectral Entropy	RMS	Pitch
	Spectral Flatness	Energy	
	Spectral Kurtosis		
	Spectral Skewness		
	Spectral Spread		
	Spectral Centroid		
	Spectral RollOff		
	MFCC		

Table 3.2: Audio descriptors adopted in the content-based analysis in the Emo-Soundscapes

Table 3.2 presents the set of audio descriptors included in the Emo-Soundscapes. Following Peeters [51], the descriptors are organised into four categories: 1) temporal descriptors, computed from the audio waveform; 2) spectral shape descriptors, computed from the short time Fourier transform of the signal, 3) energy descriptors, referring to the energy content of the signal and 4) harmonic descriptors, computed from the sinusoidal harmonic modelling of the signal.

For each soundscape in the Emo-Soundscapes dataset, a unique global descriptor vector is provided, which overall includes a single value per descriptor. To this end, descriptors are first computed on instantaneous and overlapping analysis windows across the entire 6-second duration of each soundscape. Then, to provide a unique value per description, the instantaneous analysis per descriptor is averaged.

3.3 Mapping Affective Annotations to Low-level Audio Descriptions

In light of our aim to establish a mapping between affective annotations and low-level audio descriptions, we first quantized the affective 2-dimensional valence, v, and arousal, a, space into a square (11x11) matrix $M_{v,a}$. Figure 3.4 shows the quantized spaced, where all elements in the matrix $M_{v,a}$ are equidistant points in the 2-dimensional space. The quantized matrix provides a reduced and even distribution of the 1213 data points in the Emo-Soundscapes, streamlining the efficacy (i.e., computational cost) of further processing.

For each matrix index, we then computed two vectors. The first is a descriptor vector including all descriptors listed in Table 3.2, whose computation we detail in Section 3.3.1. The second is a ranking vector of the relevancy of each descriptor, whose computation is detailed in Section 3.3.2.

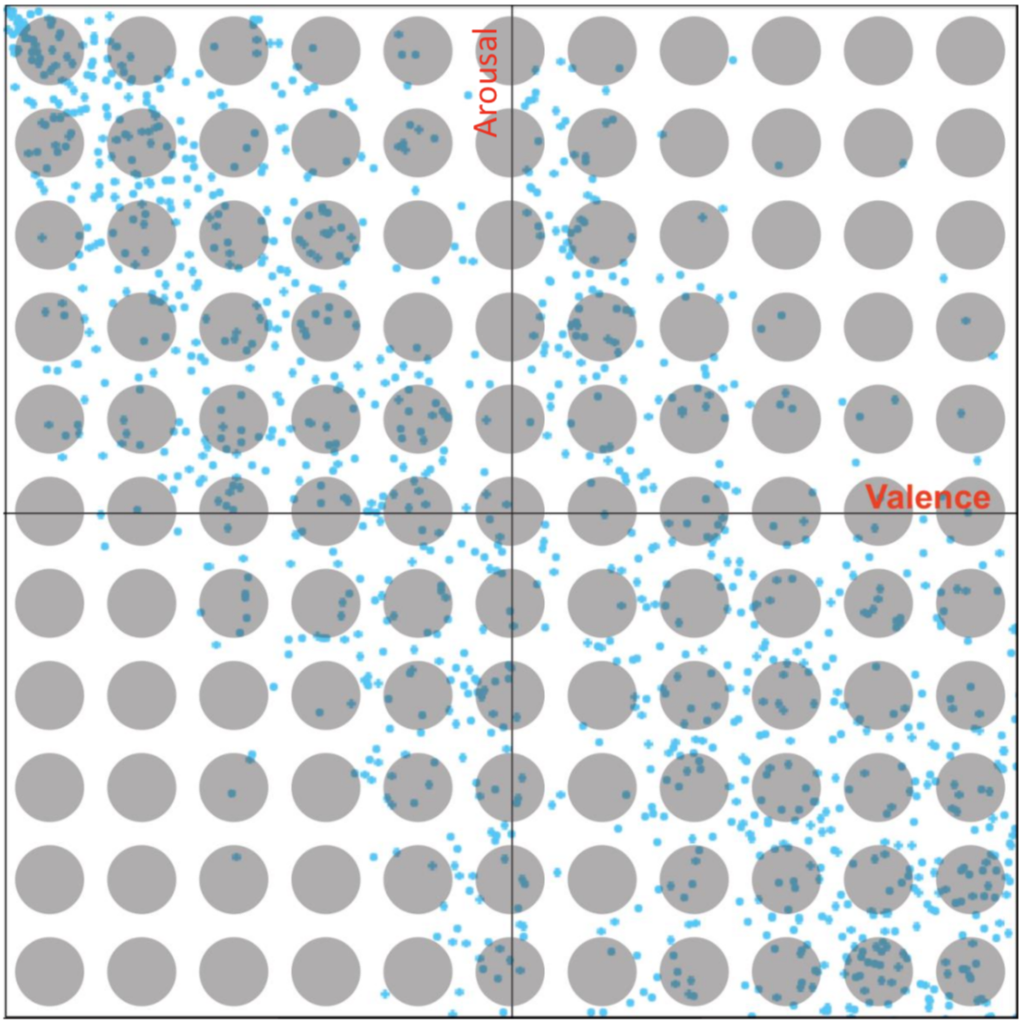


Figure 3.4: Overlap between the affective annotations of the Emo-Soundscapes in the 2-dimensional plane and the equidistant of the space into a (11x11) matrix, $M_{v,a}$.

3.3.1 Low-level Audio Descriptor Vector of Affect

For each matrix index $M_{v,a}$ a low-level audio descriptor vector was computed. It results from the mean values per descriptor of a set of k-nearest neighbour points to the center of the matrix index location in the 2-dimensional affective space. In total, $(11*11) = 121$ vectors were created. We adopted $k = 5$ nearest neighbours as a representative number of neighbour points in the less dense areas of the plane. Figure 3.5 presents the calculation of $M_{0,0}$ index description vector. Through this calculation, we aim to obtain a descriptor capable of representing a part of the affective plane. This descriptor is often called fingerprint descriptor. The fingerprint audio descriptor that will be tested as a technique to recover audio recordings with an affective charge from web audio archives.

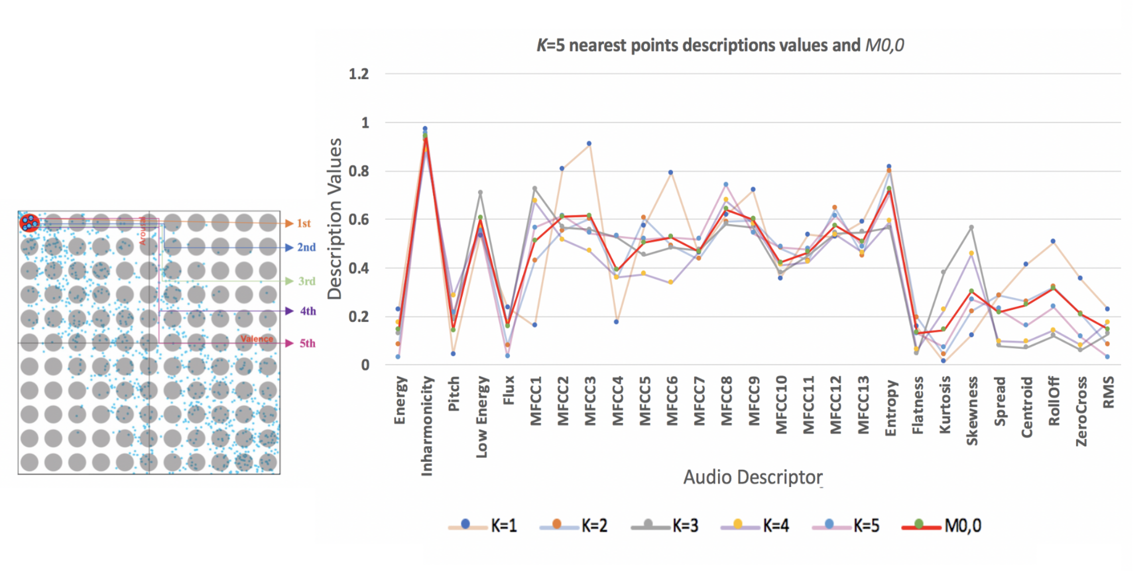


Figure 3.5: Calculation Example of a Quantized Point - $M_{0,0}$.

Through this calculation, we obtain a descriptor composed of several descriptors whose values change along the affective plane. This descriptor is often called fingerprint descriptor. The fingerprint audio descriptor will be tested as a technique to recover audio recordings with an affective charge from web audio archives.

3.3.2 Relevancy of Audio Descriptors as Ranked Variance

In line with previous studies in content-based analysis [52, 43], which typically adopt dimensionality reduction techniques to decrease the amount of processed data as well as the redundancy in the audio descriptors, we provide a ranking of the descriptor data per matrix index, $M_{v,a}$. We follow [10] and [6] in ranking descriptors relevancy based on their variance across the $k = 5$ nearest neighbours. The lower the descriptor variance, the more relevant it is for a given matrix index, $M_{v,a}$, as it shows the descriptor consistency across all k-nearest neighbours. Ultimately, the resulting ranking might provide a better understanding of the relevancy of the descriptors in the affective space, which can then enhance the retrieval of soundscapes from the circumplex affective model.

To illustrate the descriptors relevancy in the matrix $M_{v,a}$, we show in Figure 3.6 the mean variances of descriptors of the indexes $M_{0,0}$,$M_{1,0}$ and $M_{0,1}$. In the Figure 3.6, the barplots present the results from the less relevant descriptor to the the most relevant descriptor (left to right). To this purpose, we computed the mean variance of all MFCC bands in order to get a single value of variance. By ranking audio descriptors relevancy according to their variance, we verify that there is a clear tendency for the most relevant set of descriptors to be the same in these three indexes. Even though the order of relevancy of the descriptors is identical between the three points, there is an overwhelming discrepancy in the amount of variance between $M_{0,0}$ and $M_{1,0}$ with $M_{0,1}$. However, this discrepancy is more prominent among the less relevant descriptors, which is interesting to note in our study. Full relevancy analysis can be seen in A.1.

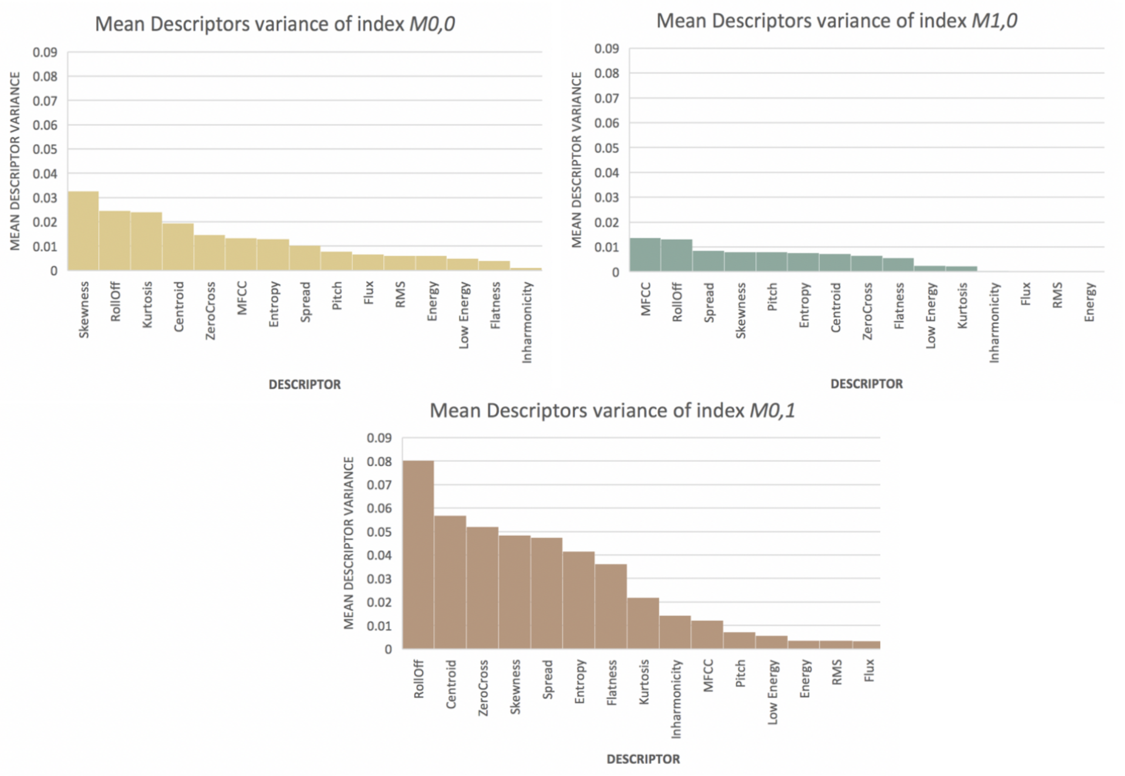


Figure 3.6: Mean Descriptors Variance: Indexes $M_{0,0}$,$M_{1,0}$ and $M_{0,1}$

Our system adopts the strategy of depicting the relevance of the descriptors through the analysis of their variances . In light of our aims, with these analysis, we intent to grasp if a set of descriptors can best capture perceived affective states. With this analysis, we were able to observe that there is in fact a set of more relevant descriptors. Even so, there is a need to understand whether this trend is recurrent throughout the affective plane, as well as quantifying the number of descriptors which can be considered the best set of descriptions.

3.4 Mining the Mappings of Affective Annotations to Low-level Audio Descriptors

In this section we provide an analysis of the descriptors variance (i.e., relevancy) per quadrant and expose the correlation of the descriptor vectors across the $M_{v,a}$ matrix dense diagonal. From the resulting patterns, we aim to understand if there is a descriptor set that best captures the perceived affect of soundscape sounds and if different affective states are expressed differently (i.e., using a clear range) in the adopted descriptors.

Each of the barplots in Figure 3.7 presents the audio descriptors from left to right according to their degree of variance. We shall highlight that the resulting variance from the second and third quadrants might be less significant due to their reduced density (please refer to Figure 3.4).

The ranked lists of variance are particularly homogeneous across all quadrants of the affective plane, notably in the five (best ranked) descriptors with lower variance. For example, the spectral descriptor *Energy, Spectral Flux and RMS* attain the lowest variance in all four quadrants. Together with this set of descriptions, *MFCC, Pitch and Spectral Kurtosis* are steadily present in the group of spectral descriptions with lower variance. This uniformity suggests that these descriptors are of greater relevancy to this set of soundscape sounds in capturing their affective states. In Chapter 5.1, we further this consideration by testing the soundness of our model using a reduced set of five audio descriptors in capturing the affective states.

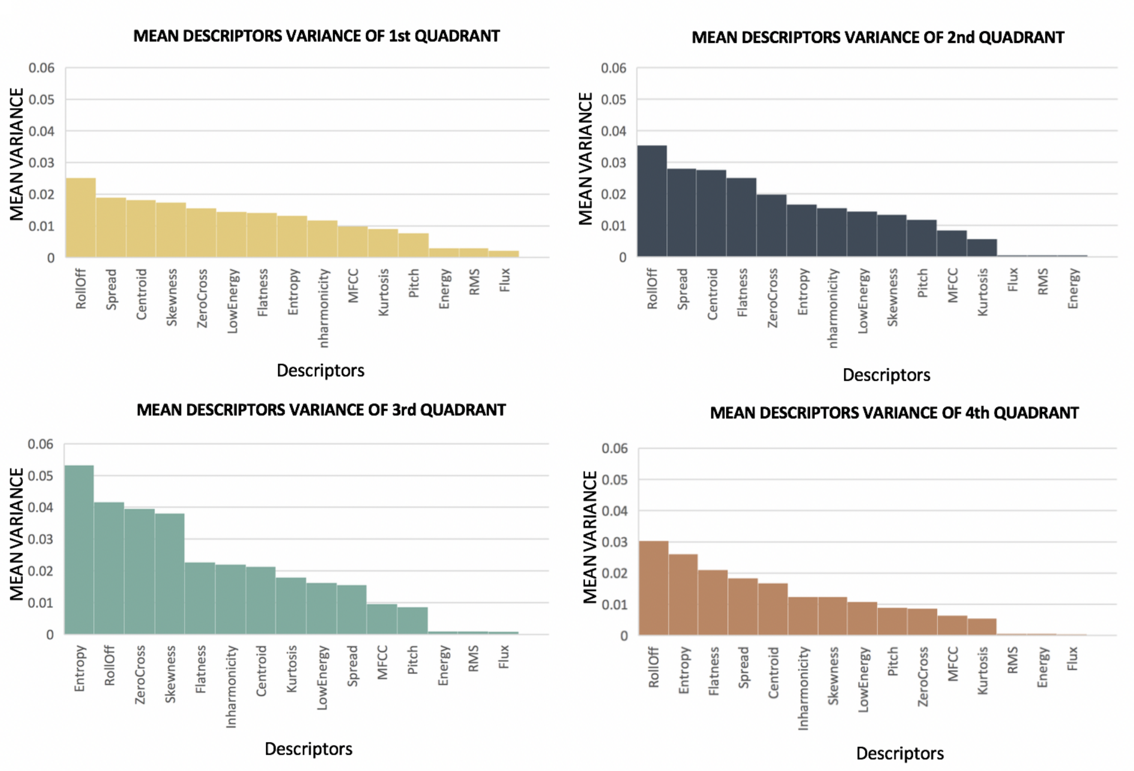
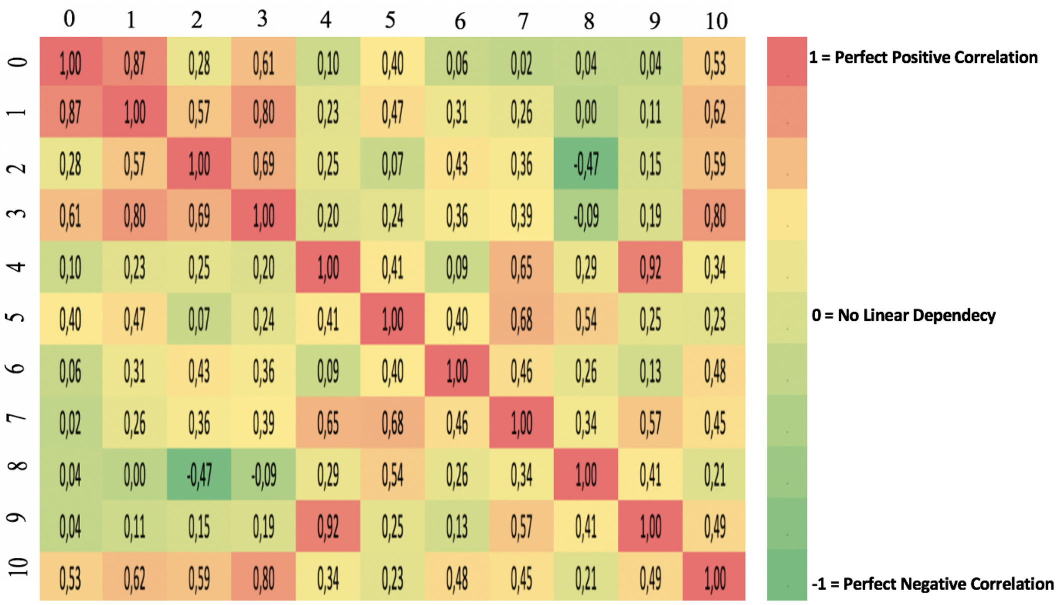


Figure 3.7: Mean Descriptors Variance of all Valence and Arousal quadrants

To supplement the view of the descriptors, in Figure 3.8, we show a self-similarity matrix that inspects the degree of correlation of the descriptor vectors across the most dense diagonal of the matrix $M_{v,a}$, i.e, the diagonal that traverses the indexes from $M_{0,0}$ to $M_{10,10}$. We adopt the Pearson correlation coefficient to measure the linear correlation between the descriptors vectors.



	0	1	2	3	4	5	6	7	8	9	10
0	1,00	0,87	0,28	0,61	0,10	0,40	0,06	0,02	0,04	0,04	0,53
1	0,87	1,00	0,57	0,80	0,23	0,47	0,31	0,26	0,00	0,11	0,62
2	0,28	0,57	1,00	0,69	0,25	0,07	0,43	0,36	-0,47	0,15	0,59
3	0,61	0,80	0,69	1,00	0,20	0,24	0,36	0,39	-0,09	0,19	0,80
4	0,10	0,23	0,25	0,20	1,00	0,41	0,09	0,65	0,29	0,92	0,34
5	0,40	0,47	0,07	0,24	0,41	1,00	0,40	0,68	0,54	0,25	0,23
6	0,06	0,31	0,43	0,36	0,09	0,40	1,00	0,46	0,26	0,13	0,48
7	0,02	0,26	0,36	0,39	0,65	0,68	0,46	1,00	0,34	0,57	0,45
8	0,04	0,00	-0,47	-0,09	0,29	0,54	0,26	0,34	1,00	0,41	0,21
9	0,04	0,11	0,15	0,19	0,92	0,25	0,13	0,57	0,41	1,00	0,49
10	0,53	0,62	0,59	0,80	0,34	0,23	0,48	0,45	0,21	0,49	1,00

Figure 3.8: Self-similarity matrix of the diagonal matrix points $M_{0,0}$ to $M_{10,10}$

We can easily identify that there are significant differences nearly everywhere in the diagonal matrix indexes, $M_{0,0}$ to $M_{10,10}$. In particular, there seems to be less correlation between most of the matrix indexes and the first, $M_{0,0}$. In addition, the $M_{9,9}$ and $M_{8,8}$, also indicates low correlation with the indexes $M_{0,0}$, $M_{1,1}$, $M_{2,2}$ and $M_{3,3}$. Although other occasional dissimilarities occur, these two occurrences are the most significant given its apparent low correlation with large part of the quantized space. On the other hand, some indexes seem to provide compelling high correlation results in considerable part of the space. For example, the $M_{3,3}$ and $M_{9,9}$ show high correlation with the indexes $M_{0,0}$, $M_{1,1}$ and $M_{2,2}$. In short, the relevance of descriptors must be attained by calculating the rankings, based on variance as seen above, per each matrix index.

3.5 Summary

In this Chapter, we presented a strategy to map affective states between perceived affective dimensions and audio descriptors in soundscapes sounds. To this end, we first exposed the challenges of assessing affective experiences, due to their ambiguous and intercorrelated nature, which motivated the adoption of a two-dimensional model of affect, namely Russel's 2-dimensional circumplex model [55].

We then presented in Section 3.2 Emo-Soundscapes, an annotated dataset for soundscape emotion recognition. We described the Emo-Soundscapes project from: 1) the collection of audio recordings, 2) the management of the dataset and 3) the affective annotations of each audio recording. Following our considerations on the distribution of the dataset affective annotations in the 2-dimensional valence-arousal space, we conclude by providing an overview of annotation space, namely the high density in the continuum that ranges from tense to calm.

We additionally provided a mapping strategy between affective annotations and low-level audio descriptors, using a quantized affective space. This procedure allowed us to extract for each quantized point of the space a representative descriptor vector of affective state as well as a ranking vector of the relevancy of the descriptors. We observed that *Energy, Spectral Flux* and *RMS* are the audio descriptors that often attain lower variance in the affective space. Ultimately, to better analyse the descriptors, we computed a self-similarity between the diagonal matrix indexes $M_{0,0}$ to $M_{10,10}$. The resulting correlation matrix exposes a lack of correlation across the diagonal matrix descriptor vectors. Hence, we can conclude that distinctive vector for each of the diagonal matrix indexes exists.

Chapter 4

MScaper

This chapter details MScaper, a system for retrieving soundscape sounds from the online crowd-sourced web audio archive Freesound. We aim to foster a fluid and streamlined soundscape retrieval method to support the creation of soundscapes using large web audio archives. Russel's circumplex model, namely the description of affect based on the combination of valence and arousal dimensions, is adopted to filter textual soundscapes queries from Freesound. In Section 4.1, we present the architecture of the MScaper system. In Section 4.2, we detail the communication with the Freesound database, namely for querying and retrieving soundscape sounds. In Section 4.3, we describe how the affective model is integrated in MScaper. In Section 4.4, we describe the capability of MScaper to adapt to a particular user preference. Finally, in Section 4.5, we describe the integration of MScaper system in Ableton Live and its interface.

4.1 Architecture of MScaper

MScaper was developed in Max MSP considering the easy prototyping and its easy integration into audio production environments. Mostly, it's compatibility with Ableton Live Digital Audio Workstation (DAW) as M4l device is ideal to present MScaper to the expert-users inside an environment which is familiar to them. Figure 4.1 shows the architecture of MScaper and the flow of information between its main component modules. MScaper is endowed of automatisms, the users have only two essential methods for interacting with the interface: tags and a navigable 2-dimensional affective plane.

The architecture of MScaper has three core components in its backend: 1) the affective valence and arousal model, 2) the Fs.Library and 3) assistive generation. The affective valence and arousal model consists in the construction in MScaper of the affective quantized space, as showed in previous chapter, as well as the implementation of the algorithms for ranking audio descriptors. Fs.library is a collection of functionalities for communicating with Freesound archive through their API services. Ultimately, the assistive generation module is the deployment of mechanisms to aid expert-users to compose soundscapes using Freesound.

The output of the system is a ranked set of soundscape audio recordings, which can be easily accessible for the design of sonic environments in audio production environments.

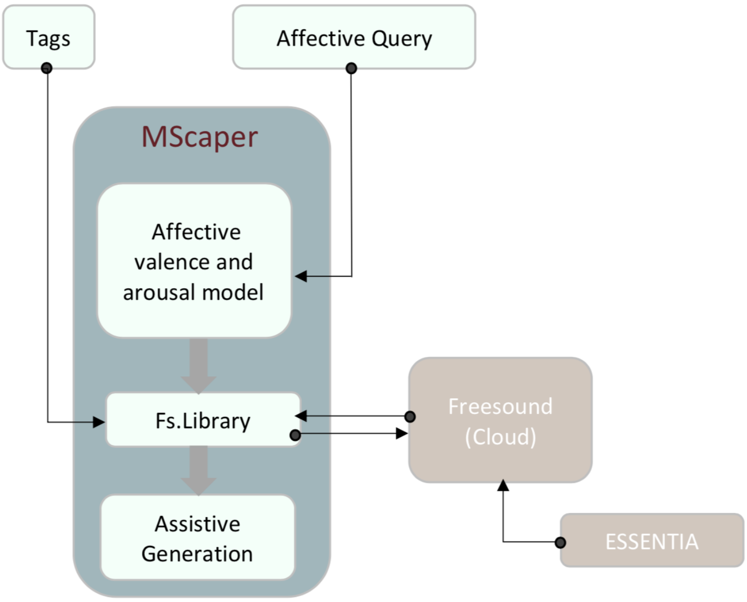


Figure 4.1: MScaper Architecture

4.2 Fs.Library

The Fs.Library is an instrumental module in MScaper. It is responsible for querying, retrieving and previewing audio recordings from the Freesound web audio archive within Max[1]. A client/server scheme is adopted. Fs.Library is the client and the Freesound the server. The library consists of the following set of five Max abstractions[2]: `fs.urlparams`, `fs.request`, `fs.retrieve`, `fs.preview` and `fs.download`.

User-defined queries (e.g., semantic tags and audio descriptions) defined in Max as messages[3] are then formatted to query the Freesound server using its dedicated API[4]. On the other hand, the Freesound server returns a variable number (ranked) soundscapes matching a given query. Ultimately, these abstractions allow sound designers to use Freesound web archives within Max,

[1] Fs.Library can be download in:
https://sites.google.com/view/soundscapeusingwebaudioarchive/home

[2] In Max, abstractions are *patches* that can operate in separate. A reusable file that you can then use inside of any patcher.

[3] *Message* object displays and sends any given message with the capability to handle specified arguments.

[4] https://freesound.org/docs/api/overview.html accessed on June 27th, 2019

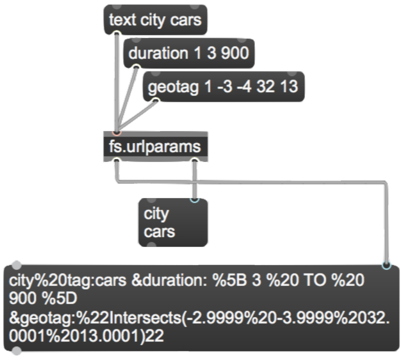
and consequently, in audio production environments such as Ableton Live, which integrates the MScaper and the Fs.Library as a Max4Live device[5].

By using Freesound API, it is possible to browse, search and retrieve data about Freesound's audio files, users and packages. Plus, to find audio recordings that are similar to a given target through content analysis and also retrieve descriptions from audio files using the Essentia library. This library contains a set of algorithms that implement audio input/output functionality, standard digital signal processing blocks, filters, algorithms for statistical data characterisation, and a large set of spectral, temporal, tonal, and high-level music descriptors.

Each of the Fs.Library abstractions are detailed at length next.

4.2.1 fs.urlparams

The `fs.urlparams` abstraction from the Fs.Library formats a message using the HTTP protocol to query the Freesound server. The `fs.urlparams` abstraction is able to sequentially collect different attributes to be merged into a single HTTP request. This information is defined by the user using message box in Max (see Figure 4.2), and can be understood as filters that narrow the Freesound web audio search space.



```
text city cars
duration 1 3 900
geotag 1 -3 -4 32 13
fs.urlparams
city
cars
city%20tag:cars &duration: %5B 3 %20 TO %20
900 %5D
&geotag:%22Intersects(-2.9999%20-3.9999%2032.
0001%2013.0001)22
```

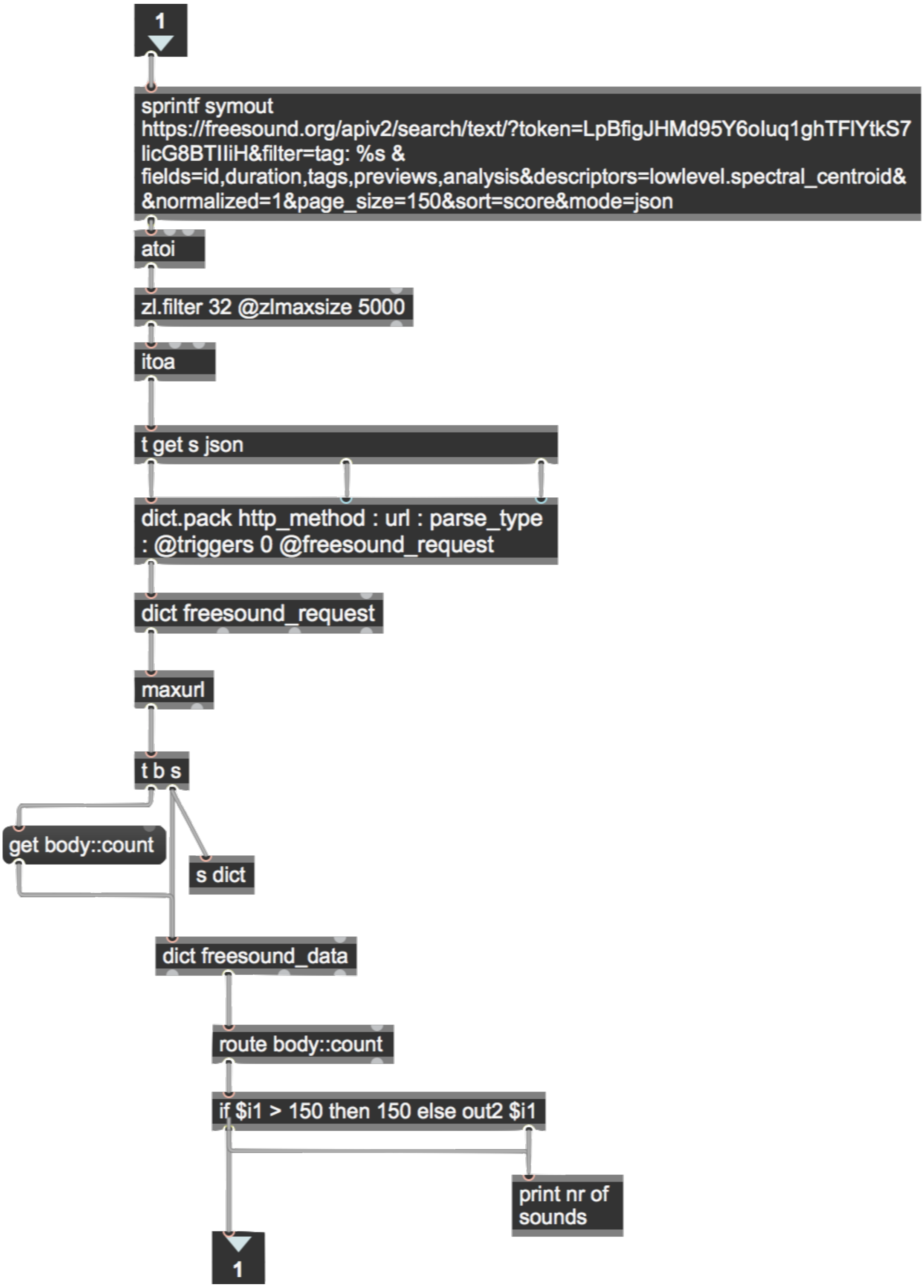
Figure 4.2: fs.urlparams abstraction

To request audio recordings with durations between 3 and 900 seconds, the user needs send to `fs.urlparams` abstraction the following message: *duration 1 3 900*. The first element in each message defines the type of filter to be applied. The remaining elements vary according to the type of filter. In the aforementioned example, the second element is a boolean data type to activate and deactivate the filter followed by two minimum and maximum parameters (in seconds) 4.2. The HTTP request is output only after a text query, i.e. textual tags, is input. In its current version, the `fs.urlparams` abstraction can filter content based on duration, geolocation and tags. The output of this abstraction are the introduced parameters formatted to join the HTTP request. To this end, it is mandatory to `fs.urlparams` be connected with `fs.request`.

[5] Max for Live device is a Max patch that can be integrated within Ableton Live environment.

4.2.2 fs.request

The `fs.request` abstraction allows to search for audio recordings by tags and low-level audio descriptors and by similarity to a given audio file in the Freesound archive. The `fs.request` is structured in three main cores: 1) the transformation of formatted strings from fs.urlparams into an HTTP request string, 2) the request to the API, and 3) the storage of incoming data from Freesound. We use the *sprintf* object (similar to the C programming syntax) to concatenate the sequential parameters of fs.urlparams abstraction into a single HTTP request. Then, a request to the Freesound API is made using the *maxurl* object. Once the request is done, the *maxurl* returns a dictionary of information containing the results of our request using the JSON format. Audio content-based queries to Freesound are outputted as a ranked list. The ranking order is computed as the Euclidean distance to the string of descriptors in the query. Thus, the sounds each dictionary are sorted by distance. The outlet of this `fs.request` sends the number of audio recordings available to access.



```
1
sprintf symout
https://freesound.org/apiv2/search/text/?token=LpBfigJHMd95Y6oluq1ghTFIYtkS7
licG8BTIIiH&filter=tag: %s &
fields=id,duration,tags,previews,analysis&descriptors=lowlevel.spectral_centroid&
&normalized=1&page_size=150&sort=score&mode=json
atoi
zl.filter 32 @zlmaxsize 5000
itoa
t get s json
dict.pack http_method : url : parse_type
: @triggers 0 @freesound_request
dict freesound_request
maxurl
t b s
get body::count
s dict
dict freesound_data
route body::count
if $i1 > 150 then 150 else out2 $i1
print nr of
sounds
1
```

Figure 4.3: fs.request abstraction structure

4.2.3 fs.retrieve

The fs.retrieve abstraction accesses the output Freesound data collected in the dictionary from the server. It allows users to retrieve specific information by navigating in the several layers of the dictionary, namely: 1) tags; 2) sound identifiers (ID); 3) preview links; and 4) audio content analysis (i.e., descriptors).

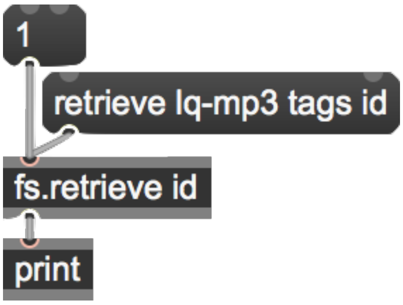
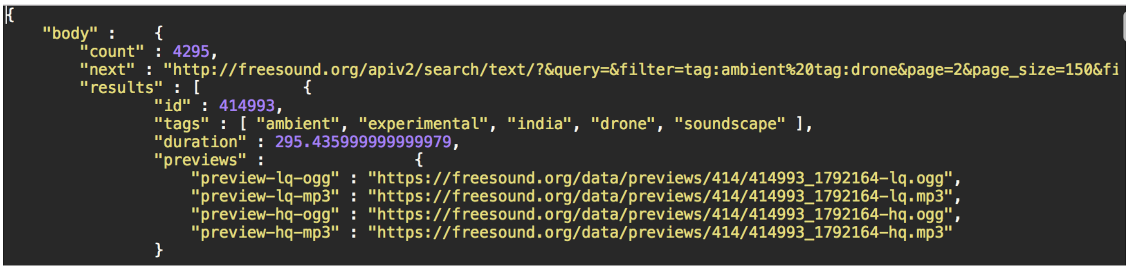


Figure 4.4: fs.retrieve abstraction

To retrieve data from the dictionary, the user must send a message to the fs.retrieve abstraction starting with *retrieve*, followed by a pointer of the content retrieved (see Figure 4.4). Then, a message with the number of the audio recording index in the dictionary will make the abstraction to output the retrieved data concerning the specified audio recording. The desired data to retrieve can also be define as arguments of the abstraction. For example, in Figure 4.4, the fs.retrieve abstraction has the argument id which makes it output the sound identifier of the sound from the dictionary user specifies.

As aforementioned, the dictionary of results is returned in JSON format. This is a file format that is used to store data in a set of key-value pairs. The Figure 4.5 shows an example of a JSON string.



```
{
    "body" :    {
        "count" : 4295,
        "next" : "http://freesound.org/apiv2/search/text/?&query=&filter=tag:ambient%20tag:drone&page=2&page_size=150&fi
        "results" : [           {
                "id" : 414993,
                "tags" : [ "ambient", "experimental", "india", "drone", "soundscape" ],
                "duration" : 295.43599999999979,
                "previews" :                {
                    "preview-lq-ogg" : "https://freesound.org/data/previews/414/414993_1792164-lq.ogg",
                    "preview-lq-mp3" : "https://freesound.org/data/previews/414/414993_1792164-lq.mp3",
                    "preview-hq-ogg" : "https://freesound.org/data/previews/414/414993_1792164-hq.ogg",
                    "preview-hq-mp3" : "https://freesound.org/data/previews/414/414993_1792164-hq.mp3"
                }
```

Figure 4.5: Example: Dictionary of Results in JSON format

From this example, it is possible to see that keys are wrapped in double quotes, a colon separates the key and the value, and the value can be of different types of data (e.g. string, integers, floats etc). When using JSON, to parse information from dictionaries it is needed to request them with specific formatting. For example, to get the audio preview link in high quality MP3, we need to reach the various levels of the JSON file with the request `get body::results[sound index (this case: 1)]::previews::preview-hq-mp`. The output of the system would be `https://freesound.org/data/previews/414/414993_1792164-hq.mp3`.

4.2.4 fs.preview

The `fs.preview` abstraction retrieves from the the Freesound server a low-quality version of an audio file. To request a given audio file the user must specify an audio identifier (ID). The abstraction is then capable of playing the audio file using the *jweb* max object (see Figure 4.6).

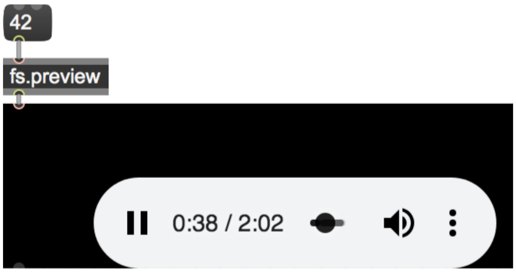


Figure 4.6: fs.preview abstraction

4.2.5 fs.download

The fs.donwload downloads audio recordings from Freesound web archive to the local disk. The file type can be defined by the user between low and high quality MP3 files (for 64kbps quality and 128kbps) and low and high quality Ogg files (for 80kbps quality and 192kbps). The process for download is made defining the filename in the second inlet of the abstraction and then specifying the position of a sound in the dictionary list.

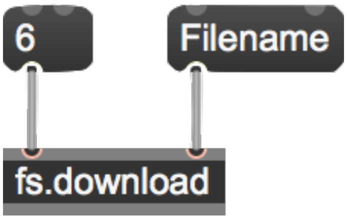


Figure 4.7: fs.download abstraction

4.3 Affective queries to Freesound via Audio Descriptors

The affective queries to Freesound within MScaper are made via the content-based filtering method in the Fs.Library. In detail, affective states are defined as a sound descriptor vector query, following the strategy detailed in Chapter 3. However, small discrepancies in the descriptor data from the Emo-Soundscapes and Freesound exist. This discrepancy results from two factors. First, they adopted different libraries to compute content-based descriptions. The Emo-Soundscapes dataset adopted the MIRToolbox [38] and YAAFE [20], while the Freesound adopted Essentia [9]. Second, the soundscapes from Emo-Soundscape dataset were retrieved from Freesound and trimmed to 6-seconds long. The adoption of a global descriptors per file might have imposed a degree of change over the descriptor analysis. To mitigate the disconnect, we first measured their correlation and applied further processing.

Per adopted descriptor, we computed the Pearson correlation coefficient across the 614 6-second unmixed sounds descriptions between the Emo-Soundscapes and Freesound. Figure 4.9, shows the resulting correlation coefficients. High correlation can be interpreted as the result of descriptors which is less prone to temporal changes. We retained all descriptors whose correlation was above a 0.4 threshold value. The dashed orange line in Figure 4.9 shows the threshold cut. In light of noticeable different across the different MFCC coefficients together with the impossibility to query Freesound using individual MFCC componenets, we discarded this descriptors. Table 4.1 presents the final set of descriptors adopted in MScaper. To minimise the use of different libraries in audio-content description, we further scale the Emo-Soundscapes descriptors to better match the Essentia descriptors. The average value across ratio differences per descriptor was adopted as the scaling factor.

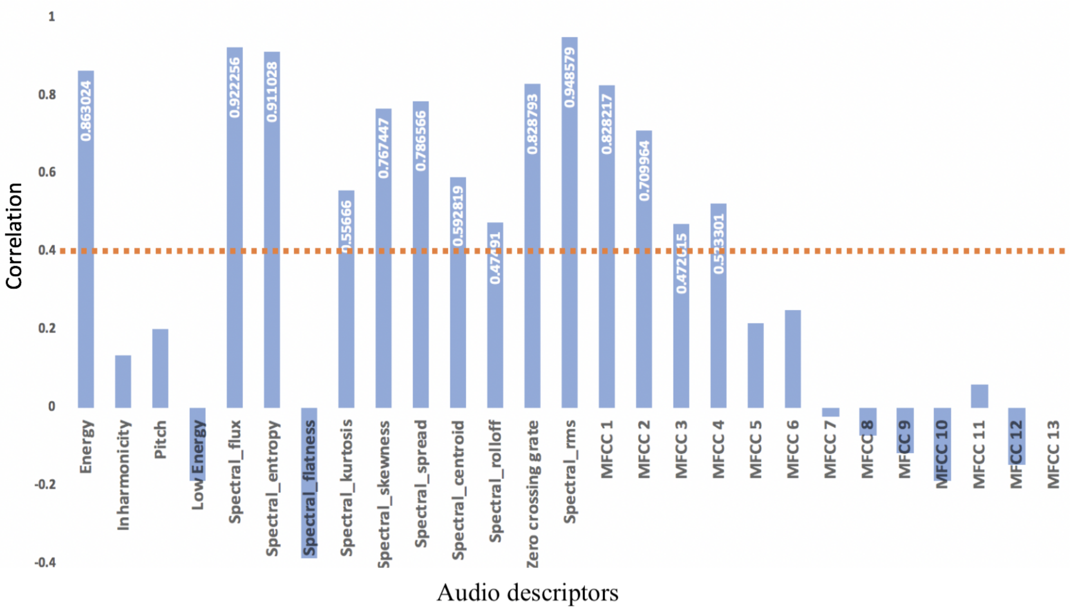


Figure 4.8: Vertical Correlation between Descriptors of Emo-Soundscapes and Freesound

Valid Audio Descriptors	
Energy	Spectral Flux
Spectral Entropy	Spectral Kurtosis
Spectral Skewness	Spectral Spread
Spectral Centroid	Spectral RollOff
Zero crossing rate	RMS

Table 4.1: Best Descriptors Based on Vertical Correlation between Emo-Soundscapes and Essentia

User-based queries of textual and content-based descriptions typically reduce substantially the results in the database. To guarantee the number of user-requested audio recordings from Freesound, we implemented a strategy that iteratively discards descriptors from the descriptor vector until retrieving the requested number of files. By default, MScaper requests Freesound 20 audio recordings per query and expects to receive the minimum of 15. Therefore, the system has integrated a browsing refinement mechanism that is activated once a query does not retrieve the minimum of 15 audio recordings. Following the strategy defined in Section 3.3.2, we discard sequentially the descriptors from the descriptor vector by relevancy. In other words, we discard sequentially the descriptors with higher variance.

4.4 User-driven annotations

Following the discussion in Section 3.1 on the subjectivity of the affective experiences, we adopt a strategy to fine-tune and personalise the MScaper system to meet the preferences of a particular user by continuously reinforcing the affective annotation space. To this end, we collect and map the audio descriptions of chosen files to user-defined affective queries at runtime. Hence, MScaper updates over time the affective model, which initially relies solely on the crowdsourced affective annotations from Emo-Soundscapes.

In greater detail, we developed a mechanism that adapts MScaper to to user preferences using their own queries over time. Figure 4.9 presents the two-dimensional valence-arousal affective plane in MScaper. The blue circle is the query location in the plane, i.e., a query for a high valence and high arousal audio recording. MScaper will present the ranked results according to the query. If a sound recording is selected, its audio descriptor is then mapped to that particular affective location in the space. Thus, updating the Emo-Soundscape annotations. Furthermore, users can even provide further refinements. By pressing the *user* icon top-left side of the plane a corrected yellow cross is added to the plane, meaning that the system entered a learning mode. In Figure 4.9, the user has corrected its selected file affect to a state closer to bored or sad. To achieve the reinforcement, users need to drag the blue circle to the desired point in the space.

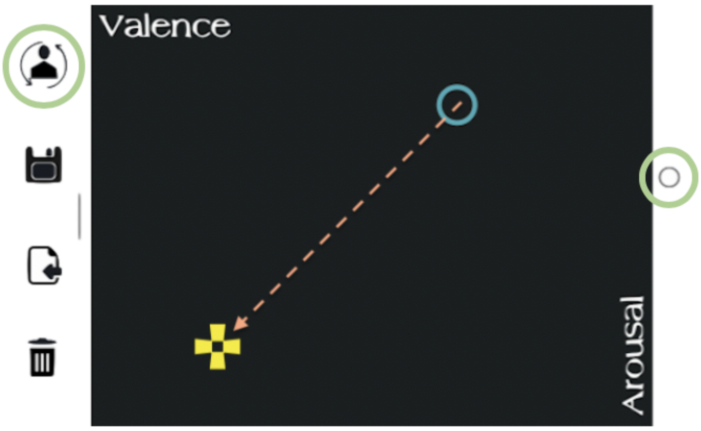
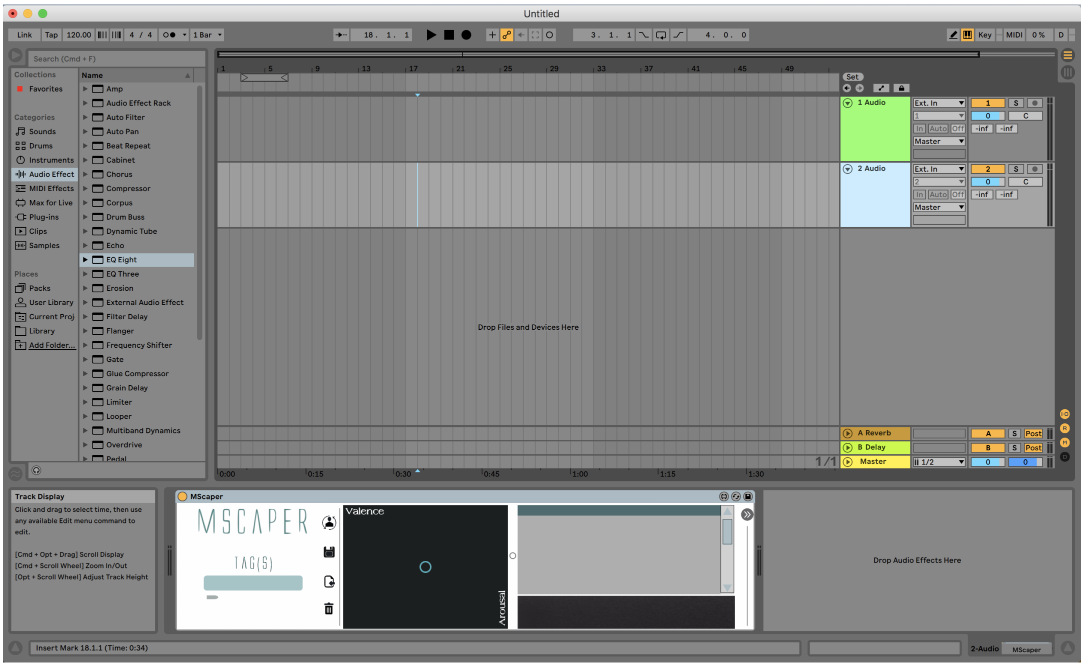


Figure 4.9: Remapping the Affective Model with User-driven Affective Annotations

4.5 Interface and Integration in Ableton Live

Figure 4.10 shows the MScaper (Max4Live) interface within the Ableton Live DAW. It is divided into three main parts (from left to right): 1) the query data, where users formulate build the requests composed of textual tags and a point in the affective plane, 2) the results table, where users access and preview the retrieved set of soundscape audio recordings and 3) the shift session module, an automatic mechanism of MScaper to progress the full set of audio recordings in the same direction.



Click and drag to select time, then use any available Edit menu command to edit.

[Cmd + Opt + Drag] Scroll Display
[Cmd + Scroll Wheel] Zoom In/Out
[Opt + Scroll Wheel] Adjust Track Height

Insert Mark 18.1.1 (Time: 0:34)

Figure 4.10: MScaper device within the Ableton Live DAW

In Figure 4.11, we present the query design part of the interface. By query design, we refer to the strategy adopted for users to input data that is then processed and sent to the Freesound to retrieve soundscapes. The interface is design so that user flow from left to right. A successful query must include a textual tag (or tags) and a position in the 2-dimensional affective space.

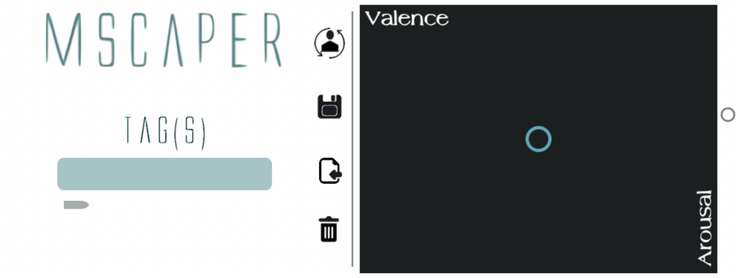
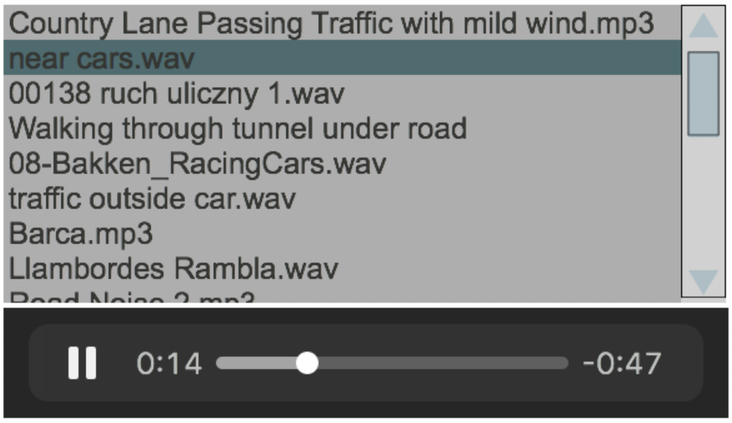


Figure 4.11: MScaper interface to formulate the queries

In the light green rectangle under the word tag(s) in Figure 4.10, users can define the textual tags according to what elements compose the soundscape. It can be defined one or multiple words, yet, a large number of tags might result in a lack of results. To ensure that the system corresponds to what is expected, users should establish tags that suits in the referential classification proposed by Schafer [62], as seen in Section 2.1. Below the green rectangle, there is a gray slider to filter the duration of results. There, users define the minimum duration of the audio recordings to retrieve, from 5 to 45 seconds long. Ultimately, the navigation in the Freesound's archive is achieved by moving the blue circle in the plane. Every time a new location is selected in the plane, MScaper queries Freesound. This is indicated by the flashing red led in the right side of plane.

Once the retrieval process is complete, the red led stops flashing, and a set of soundscape audio recordings is presented in MScaper as a ranked list. The adoption of a list relates to the familiarity of this interface design to expert-users in audio production environments. The ranking order of the results reflects the distance of the audio recording from the query. To this end, a Euclidean distance is calculated from the input query. Then, the results are ranked from the closest to the furthest. Audio recordings with the smallest distance to the query are displayed first. The Figure 4.12 presents the table of results for the tag *cars* in the extreme location of the affective plane corresponding to low valence and high arousal.



Country Lane Passing Traffic with mild wind.mp3
near cars.wav
00138 ruch uliczny 1.wav
Walking through tunnel under road
08-Bakken_RacingCars.wav
traffic outside car.wav
Barca.mp3
Llambordes Rambla.wav

Figure 4.12: Interface: The Results Table

By clicking in any audio recording name, MScaper playbacks a low quality preview of the file. The audio recordings in the list are available to users make their own assessment. Once a user finds the most suitable audio recording, s/he can double-click over the audio recording name and MScaper starts downloading the soundscape. These files are automatically stored in the Ableton's Library folder, appearing to users in the sidebar as seen in Figure 4.10.

4.6 Dynamic Session-content Manipulation of Affect

In light of the relevancy of dynamic content creation and manipulation in today's multimedia landscape, MScaper includes a mechanism that aims to experiment with affect manipulation in soundscape creative sessions. This strategy is the shift session module of the interface. To leverage the high-level transformation of affect in soundscapes, we adopt a strategy to automatically move a collection of audio recordings according to the same vector transformation in the affective plane.

Figure 4.13 shows the shift session module of the interface. In the Figure example, we transformed a tense urban soundscape into a clam urban environment. To this end, we first used MScaper to retrieve certain audio recordings from Freesound in the first quadrant of the plane. The following tags composed our soundscape: 1) people (white circle), 2) cars (orange circle) and 3) nature (purple circle). After the selection of the pretended audio recordings and download them to the Ableton's library folder, a second plane will appear by pressing the most right icon in the interface, as presented in Figure 4.10. The new interface is the transformation module. The set of audio recordings will appear around the mouse (green circle) maintaining the same distance between audio recordings as initially retrieved. Users can navigate again in this affective plane in order to create a new set of audio recordings in a different part of the plane, retaining their relative position.

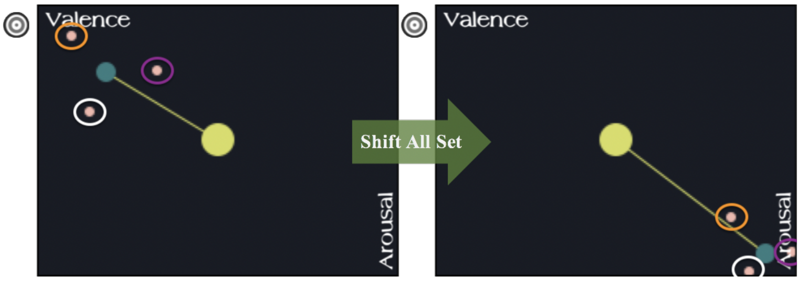


Figure 4.13: Interface: The Shift Session Module

The right image in Figure 4.13 shows the target affective state. Please note that the moment that one of the points (each audio recording) reaches the limit of the plane, the points try to adapt to the plane limits and maintain the relation between them. Ultimately, in order to browse for the target audio recordings, users must click the target icon that on the left side of the plane[6].

4.7 Summary

In this Chapter, we described the algorithms and interface of the MScaper system as well as its integration with Ableton Live. In Section 4.1, we described the MScaper architecture as well as the its different algorithms. In Section 4.2, we presented the communication with Freesound API, namely using our Max Fs.Library, which includes the following abstractions: fs.urlparams, fs.request, fs.retrieve, fs.preview and fs.download. The library provides a reliable method to enable audio experts to use the potential of the Freesound web archive within Max and Ableton Live. In Section 4.3, we detailed the audio descriptors data mining strategy used in MScaper as well as how affective queries are defined as a string of audio descriptors.

Finally, we described a browsing refinement mechanism of MScaper to query the Freesound web archive within sound production environments. Furthermore, in Section 4.4, we presented a

[6] audio recordings retrieved can be accessed in:
https://sites.google.com/view/soundscapeusingwebaudioarchive/home

strategy to reinforce the affective annotation of the MScaper system using user- or session-specific affective annotations, enabling the personalization of the system to meet the users preferences.

All the example materials mentioned in this chapter can be accessed in the following link: https://sites.google.com/view/soundscapeusingwebaudioarchive/home

Chapter 5

Evaluation

In this Chapter, we detail the evaluation of our work. Two evaluation tests are adopted. The first is a perceptual test that aims to assess the robustness of the mapping strategy between the affective dimensions and audio descriptors – detailed in Chapter 3. The second is a usability test of the MScaper prototype application – detailed in Chapter 4 – which aims to estimate the perceived usability of the tool in the context of audiovisual production environments. In Section 5.1, we detail the evaluation design of both the listening and usability tests. In Section 5.2, we present and discuss the results of the evaluation.

5.1 Evaluation Design

5.1.1 Perceptual Test

We designed a test which aims to perceptually assess if the proposed strategy to adopt low-level audio descriptors from k-nearest affective annotations in the 2-dimensional circumplex model captures soundscape affective dimensions. Ultimately, our test sheds light on the soundness of audio-content descriptors in retrieving soundscapes from large audio archives using affective dimensions. A secondary objective of our test is to understand the impact of the number of the low-level audio descriptors used in querying the dataset for capturing affective dimensions. In light of the redundancy of low-level audio descriptors and supported by the ranking established in Section 3.3.2, we compare soundscapes retrieved from audio description vectors using the entire set of adopted descriptors and a reduced number of the five best ranked descriptors per matrix index, $M_{v,a}$.

Our perceptual evaluation consisted of a online listening test with ranking questions in *SoGo-Survey*[1] survey platform. The adoption of ranking questions is due to the subjective nature of the task. This question type allows respondents to identify which objects are most and least preferred given a set of conditions. The respondents involvement was completely voluntary and anonymous. To this end, only personal information relevant to this listening test was asked.

[1] https://www.sogosurvey.com

Prior to the experiment, a training phase presented detailed instructions for the listening experiment along with a soundscape example to adjust the audio level. The participants were asked to use good quality headphones and to run the listening experiment in a noiseless environment.

Information about the expertise in sound design was collected from each participant. In particular, the level of training in sound design according to the following three categories: *No Training*, *Some Training*, and *Advanced Training*.

In total, the test included five ranking questions, each with five soundscapes sounds of 6 seconds long. The adopted duration of the soundscapes followed the suggestion of the original listening experiment by Fan et al. [23]. In the aforementioned study, the authors concluded that a 6-second duration was optimal to perceive the affect in soundscapes sounds.

Each question in the test was linked to a specific affective location in the 2-dimensional plane. Five equidistant locations along the most dense diagonal in the 2-dimensional plane (from distressed to calm) were selected. Figure 5.1 presents this selection across the diagonal. Two of the questions fall within quadrants and the remaining one is a neutral position in the center of the plane.

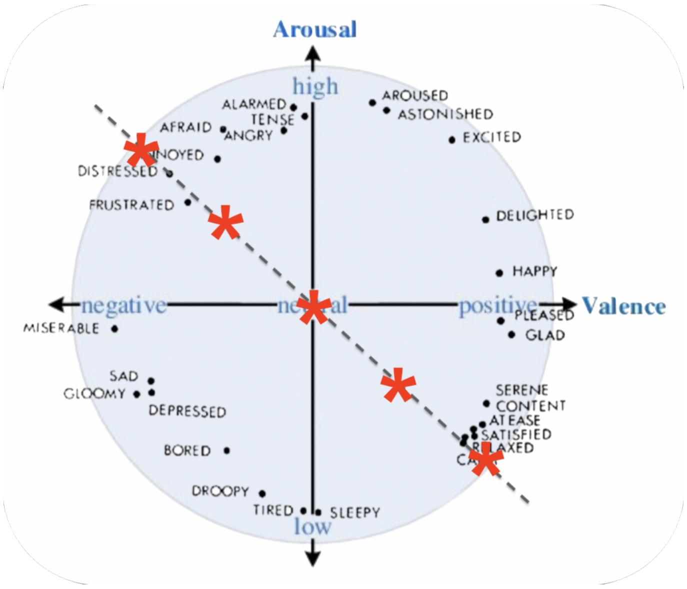


Figure 5.1: Listening Experiment: Red crosses represent the target affective state

Each of the five soundscapes per question relate to the following categories:

1. **Emo-Soundscapes**: the nearest audio recording to the $M_{v,a}$ index (See Figure 3.4);

2. **Emo-Soundscapes** *opposite*: the nearest audio recording to the opposite $M_{v,a}$ index;

3. **MScaper**: audio recording retrieved from Freesound using MScaper affective model with a vector size of ten audio descriptors;

4. **MScaper** (*using 5 descriptors*): audio recording retrieved from Freesound using MScaper affective model with a reduced vector size based on their relevancy;

5. **MScaper** *opposite*: audio recording retrieved using the descriptions of the $M_{v,a}$ index.

The aforementioned set of soundscape categories per question enable us to assess the efficacy of our affective model in retrieving soundscapes from Freesound. To this end, we compare for each of the location in the affetive space soundscapes from two main sources: Emo-Soundscapes (annotated manually) and Freesound (retrieved by low-level audio descritors). Furthermore, we include soundscapes from opposite locations of our query to test if the user rankings are effective in capturing affective states. To retrieve soundscapes from Freesound no textual tags were applied. For each specific $M_{v,a}$ index, Freesound soundscapes resulted from audio descriptor vector queries only. Manually, we retrieved the first soundscape results fitting the Schafer taxonomy [62].

Instrumental sounds featuring well-defined pitch were excluded.

Following Fan et al. [23], we trimmed the Freesound soundscapes to 6 seconds by manually identifying the most prominent event within its entire duration. To ensure a fair comparison between Freesound and Emo-Soundscapes, both were presented as mono audio soundscapes.

In order to select the opposite soundscape categories in both datasets, we retrieve a a soundscape within the maximum possible distance across all the plane locations. In other words, we consistently considered the diagonal distance within each quadrant as opposite location, since this distance could be guaranteed for all locations. Figure 5.2 shows the opposite location (orange dot) of a an affect in the plane (marked as a red cross).

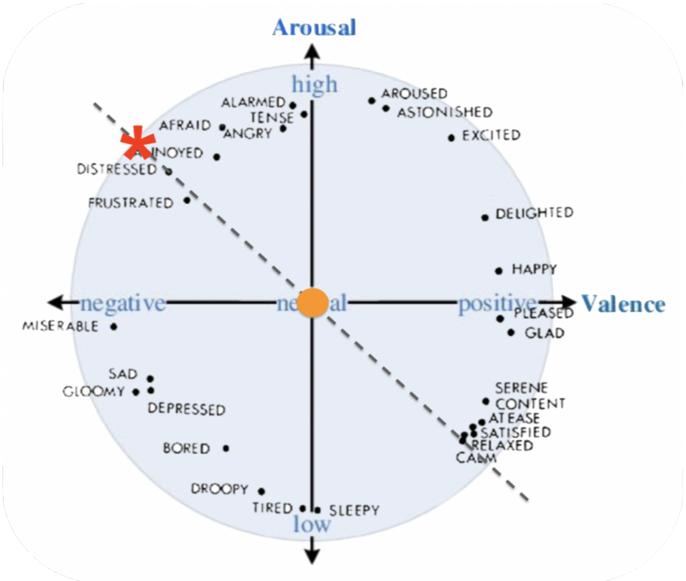


Figure 5.2: The opposite index (orange circle) in the plane in consideration to the target affective state (red cross)

Participants were instructed to rank the file soundscapes per question as follow:

Please rank the following 5 audio examples according to how well it describes the emotional state indicated by the red cross in the figure below. 1 is the sound that best matches the emotion and 5 is the sound that worst matches the emotion.

Participants then proceeded by ranking each soundscape from 1 to 5. All ranking numbers had to be chosen and no repetition was allowed per question. The ranking 1 indicates the closest soundscape to an indicated affect. The ranking 5 indicates the farthest.

In order to avoid the order-effect, i.e., the influence by the order in which questions are presented in the test, we randomly sorted the questions for each participant to minimise this effect. Additionally, the audio recordings in each question are also randomly presented[2].

5.1.2 System Usability Test

We adopt a System Usability Scale (SUS) [11], originally proposed by Brooke in 1986, to measure the perceived usability of MScaper in the context of a audiovisual production environment. The SUS consists of a ten-item questionnaire with five response options ranging from 1- *Strongly agree* to 5- *Strongly disagree*. Table 5.1 presents the questions that compose the SUS test.

System Usability Scale Test Questions
1. I think that I would like to use MScaper frequently
2. I found MScaper unnecessarily complex
3. I thought MScaper was easy to use
4. I think that I would need the support of a technical person to be able to use MScaper
5. I found the various functions in MScaper were well integrated
6. I thought there was too much inconsistency in MScaper
7. I would imagine that most people would learn to use MScaper very quickly
8. I found MScaper very cumbersome to use
9. I felt very confident using MScaper
10. I needed to learn a lot of things before I could get going with MScaper

Table 5.1: SUS test Questions

The test design was split into four main parts. First, we verify the participants expertise in sound design. Second, a brief explanation of MScaper was provided to acquaint the participants with the interface of MScaper and its integration with Ableton Live. Third, participants conducted a sound design task using MScaper. Fourth, the the participants were asked to respond to a SUS questionnaire.

Table 5.2 includes the three initial questions of the experiment, used to attest the participants' level of expertise in sound design and their level of proficiency with Ableton Live. To proceed to the MScaper evaluation, participants were required to satisfy any of the following criteria: 1)

[2] Audio recordings used in the listening experiment can be accessed in: https://sites.google.com/view/soundscapeusingwebaudioarchive/home

at least one year of professional sound design experience or 2) to be enrolled in the sound design subject from the Multimedia masters degree.

1. How do you describe your level of expertise in sound design?
2. Years of professional experience designing sound?
3. Do you usually use Ableton Live in sound design?

Table 5.2: Introductory questions of the SUS test

The experiment proceeded with a brief explanation about the outline of our research as well as the affective plane employed in MScaper. Following the explanation of MScaper interface and its integration in Ableton Live. An example of a query using MScaper (textual tag: *cars* + affective plan: *random point*) was given so they could fully conceived the practical uses of the application for generation of soundscapes. Next, an extracted video clip (15-second long) from the game "War of Rights"[3] was presented to them.

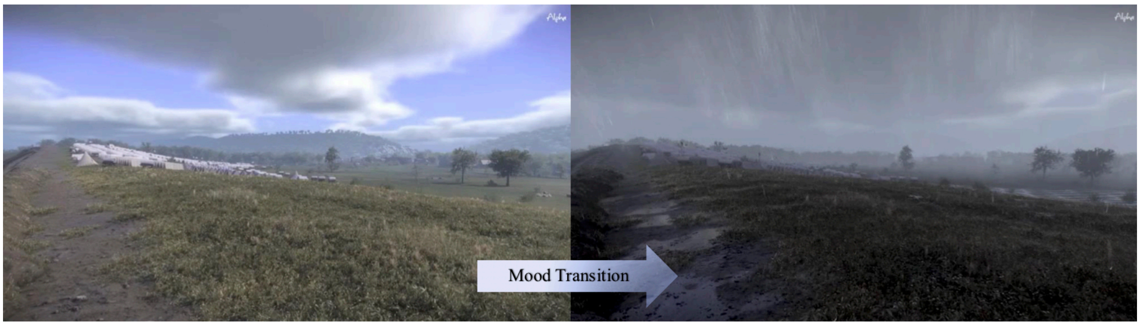


Figure 5.3: Different mood states extracted from the game *War of Rights* excerpt

The video clip excerpt form "War of Rights"[4]. displays a mountain scene where which shifts from a mild weather to a heavy storm starts. We considered the video clip particularly fit in the context of MScaper due to the noteworthy weather transition. The figure 5.3 shows two images extracted from the game that reveals the transition of mood along the excerpt. We informed all participants that the main purpose of this test was to create a soundscape, using MScaper, for this video game excerpt.

5.2 Results and Analysis

5.2.1 Perceptual Test

We collected 61 complete surveys from our online perceptual test. From the pool of participants, 44.26% (27 participants) reported having *no training in sound design*, 36.06% (22 participants) reported having *some training in sound design* and 19.68% (12 participants) reported having *advanced training in sound design*.

[3] War of Rights - https://warofrights.com/ accessed on June 30rd

[4] The video excerpt can be accessed in: https://sites.google.com/view/soundscapeusingwebaudioarchive/home

To analyse the data, we computed descriptive statistics and a paired sample t-test in *SPSS*[5] software. We adopted the descriptive statistics to provide simple summary about the sample. This gave us the first insights about the assessment results. However, in order to obtain a more in-depth data analysis, we used paired sample t-test to determine whether the results are statistically significant.

A key aim was to assess whether the differing levels of sound design experience would impact the perceived affect in soundscape audio recordings. The matter of sort out the participants according to their level of expertise in sound design showed no statistical significance when we calculated the paired sample t-test between levels of expertise.

The descriptive statistics results are presented in the boxplot of Figure 5.4. To compute the statistics we first merged all five questions per variable: 1) MScaper, 2) MScaper using 5 audio descriptions, 3) MScaper Opposite, 4) Emo-Soundscapes and 5) Emo-Soundscapes Opposite. The boxplot shows the distribution of data of all participants in the survey. The blue rectangle represents the interquartile range, the variability of the that specific audio recording and the black bold line indicates median value from the answers. Ultimately, the arrows out of the rectangles express the minimum (the lower arrow) and the maximum (the upper arrow) rank that those classes of audio recordings got. The descriptive analysis of each question can be seen in the Figure A.4.

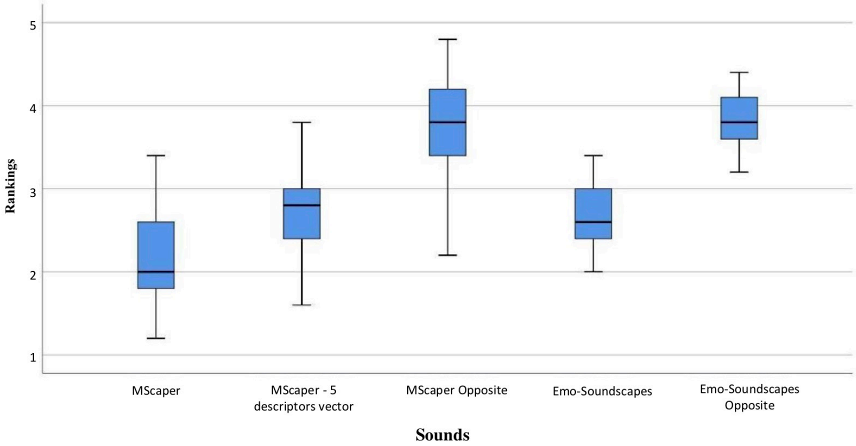


Figure 5.4: Descriptive Statistics from All Questions of Perceptual Test

The results in Figure 5.4 show that participants agreed that the soundscapes retrieved using MScaper are quite representative of the affective state under evaluation. Overall, MScaper was perceived by the participants as best matching a particular affect. Conversely, the class MScaper Opposite was also often perceived as the least representative. This descriptive statistical suggest that the MScaper, notably the descriptor vector per matrix index, $M_{v,a}$, is effective in capturing the nuances of the affective plane. The pair Emo-Soundscapes and its opposite follow the same trend. However, it is noticeable that Emo-Soundscapes results tend to be less representative than MScaper and their deviation from median is considerably lower. We believe the poorer results

[5] The IBM SPSS® software platform offers advanced statistical analysis, a vast library of machine-learning algorithms, text analysis, open-source extensibility, integration with big data and seamless deployment into applications.

of the Emo-Soundscapes are due to the smaller number of sound examples in comparison with Freesound.

Furthermore, we adopted a paired sample t-test, or a dependent sample t-test, to determine whether there are statistical differences between the variables under study in capturing affect in soundscapes. In detail, we conducted the paired t-test to assess the statistical difference between the following variable pairs: 1) MScaper / MScaper- 5 descriptors; 2) MScaper / MScaper Opposite; and 3) Emo-Soundscapes / Emo-Soundscapes Opposite. The results presented in the Table 5.3 show us that the MScaper / MScaper Opposite are the only classes statistically significant ($< .0001$).

Paired Samples Correlations

	Sample	*Correlation*	*Sig.*
MScaper / MScaper- 5 descriptors vector	61	-0.073	0.575
MScaper / MScaper Opposite	61	-0.406	0.001
Emo-Soundscapes / Emo-Soundscapes Opposite	61	-0.191	0.141

Table 5.3: Paired Sample t-test between classes of audio recordings

From the results in Table 5.3, we can conclude that the use of the entire set of descriptors or a reduced number of descriptors in the Freesound query does not seem to imply statistically significant results. Yet, the entire set of descriptors adopted indicates better overall results.

5.2.2 System Usability Scale

The SUS results are representative of a sample of ten experienced sound designers, based on a minimum assessment of having at least 1 year of professional experience, or having a degree in sound design. We invited students, and former students from the master in Multimedia, Interactive Music and Sound Design profile, from University of Porto, to participate in this usability test.

The users rated each of the 10 questions presented in table 5.1 from 1 to 5, based on their level of agreement. We then computed a SUS score per question according to the following three steps:

- For each of the odd numbered questions, subtract 1 from the score. The item's score contribution will range from 0 to 4.
- For each of the even numbered questions, subtract their value from 5. The contribution is 5 minus the scale position.
- Multiply the sum of the scores by 2.5 to obtain the overall value of SUS scores have a range of 0 to 100.

The participants agreed that MScaper was quick to learn and easy to use. Furthermore, an overall disagreement is noted on the need for further technical support or extensive training. They didn't found the tool inconvenient, inconsistent, or unnecessarily complex. However, there was only a fair agreement in wanting to use MScaper frequently. MScaper attained a mean SUS score of 82.5% with a considerable large standard deviation (STD=26.7). The table 5.4 presents the overall SUS score and the standard deviation (STD).

SUS Score	STD
82.75%	11.15

Table 5.4: SUS Score and Standard Deviation

The following Figure 5.5 presents the score of each user.

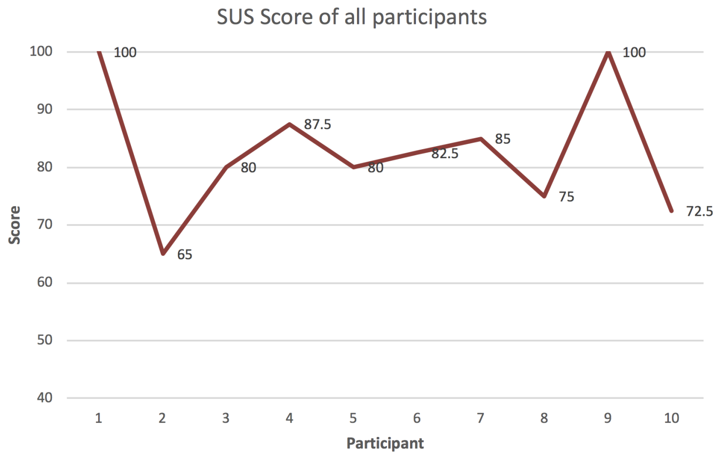


Figure 5.5: SUS Score of All Participants

MScaper overall usability attained the score of 82.75 out of 100 which is considered an excellent score regarding the perceived usability, as seen in Figure 5.6. However, the analysis with experienced Ableton Live have shown slightly improvements in the perceived usability, we do not consider this difference significant, which confirms the stability of MScaper concerning the perceived usability for overall sound designers.

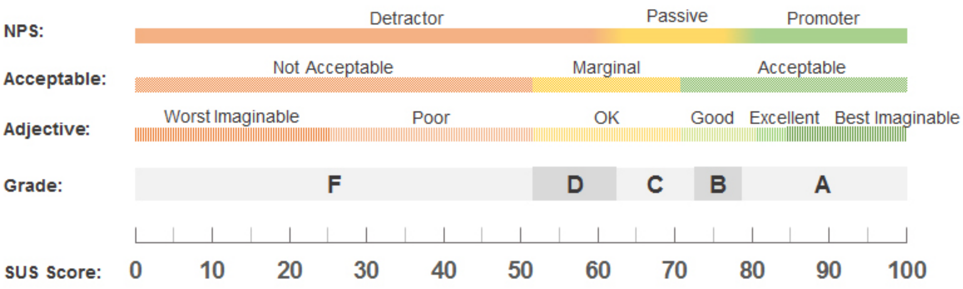


Figure 5.6: Grades, adjectives, acceptability, and net promoter score categories associated with raw SUS scores

Due to the MScaper integration in Ableton Live, we ran a test to understand the proficiency of the participants with Ableton Live to verify the impact of this variable from the SUS test. There were some participants that struggled in performing simple operations, such as cutting audio recordings or creating crossfades between them. To this end, we only considered in the SUS test the participants that reported having experience with Ableton Live above 3 (on a scale from 1 to 5). This resulted in a sample of five sound designers. The better work-flow of sound designers managing Ableton Live has shown us that is a slightly improvement in the perceived usability of MScaper. The following Table 5.5 presents the SUS score and the STD of participants with above-average workflow in Ableton Live.

SUS Score of Ableton Live Experts	STD
88.5%	11.12

Table 5.5: SUS Score and Standard Deviation of above-average Ableton Live experts

Chapter 6

Conclusions

6.1 Summary

In this dissertation, we presented MScaper, a system for navigating in the Freesound web audio archive. The system allows expert-users to retrieve soundscapes using affective (valence-arousal) dimensions. Our method expands existing technology for navigating in audio archives using a semantically-driven level in line with human perception and cognition. In detail, we established a mapping between affective dimensions and low-level audio descriptors that can be used to query the Freesound web audio archive. This strategy was studied as a fluid method for leveraging online audio commons content for multimedia production environments.

The main challenge behind our goal was to establish a mapping between affective dimensions and low-level audio features, in order to create a feature-based affective model for soundscapes. Relying on crowdsourced affective annotations from the Emo-Soundscapes, we designed and evaluated a mapping to different sets of audio descriptors. Furthermore, we ranked these descriptors by relevancy, using a dispersion or variance metric per location in the affective plane. The prototype system, MScaper, implemented our method to be tested by expert users in the context of audio production. We have integrated the prototype in the Ableton Live DAW. The results of a perceptual test evaluating our feature-based affective model and a usability test of MScaper supports the adoption of our method as a fluid mechanic for retrieving soundscapes and promoting new interfaces for sound production.

6.2 Contributions

The main novel contributions of our work can be summarized as follows:

- Fs.Library, a modular client library of the Freesound API for the Max programming environment. It allows the integration of the Freesound web audio archive in audio production environments.

- Audio feature-based affective model resulting from mapping crowdsourced affective annotations to low-level audio descriptors. It can be used to query soundscapes from web audio archives using a fluid semantically-aware audio description.

- A user-centered approach to retrieve soundscapes from web audio archives in audio production environments, namely through the integration of the MScaper system into Ableton Live, a professional audio production environment

- A dynamic strategy for soundscape manipulation and transformation within MScaper and named *Shift Session Module*, which allows users to experiment with the adaptive affect manipulation in soundscapes. This module automatically moves an entire session (i.e., a collection of previously selected soundscapes) according to a transformation vector in the affective plane.

6.3 Future Work

The MScaper system have been shown to promote novel strategies for soundscape creation by expert-users. Yet, this novel strategies shed light to a wide array of possibilities in the domains of artistic creation using web audio archives. A prime example is the use of MScaper for the automatic generation of soundscapes in light of the work of cite [67].

Moreover, the methods behind MScaper can be a great asset for the dynamic generation of audio in non-linear virtual environments, games, and interactive installations, to cite a few. The fluidity of MScaper to navigate web audio archives together with the growing array of algorithms for automatic soundscape generation can leverage the real-time dynamic creation of soundscapes in new media scenarios. In this context, we could envision adaptive soundscapes that respond to a given external or interval behaviour (e.g., user input or the mechanics of a graphic rendering system).

Our affective model results from mining collections of audio descriptors towards a ranking of their relevancy to the task at hand. In future, we aim to pursue different methods to rank audio descriptors according to their perceptual relevancy. The use of deep learning algorithms, which surpass the need for a description stage, may provide some insights on the task.

References

[1] Freesound 2: An Improved Platform for Sharing Audio Clips. *International Society for Music Information Retrieval Conference (ISMIR 2011)*, (November):4–6, 2011. Cited on pages xi, 8, and 9.

[2] Ivana Andjelkovic, Denis Parra, and John O'Donovan. Moodplay: Interactive music recommendation based on Artists' mood similarity. *International Journal of Human Computer Studies*, 2018. Cited on pages 2 and 14.

[3] De Baets, Hans De Meyer, Micheline Lesaffre, Marc Leman, and Koen Tanghe. User-dependent taxonomy of musical features as a conceptual framework for musical audio-mining technology 2, Ghent University 3 Department of Applied mathematics. *Information Systems*, 2003(SMAC 03):4–7, 2003. Cited on pages xi and 11.

[4] Emmanouil Benetos, Simon Dixon, Zhiyao Duan, and Sebastian Ewert. Automatic music transcription: An overview. *IEEE Signal Processing Magazine*, 36:20–30, 01 2019. Cited on pages 1 and 2.

[5] Gilberto Bernardes. Dynamic Music Generation, Audio Analysis-Synthesis Methods. (March), 2019. Cited on page 2.

[6] Gilberto Bernardes, Luis Aly, and Matthew E. P. Davies. SEED: Resynthesizing Environmental Sounds from Examples. *Proceedings of the Sound and Music Computing Conference (SMC)*, (January):55–62, 2016. Cited on pages 2 and 26.

[7] Gilberto Bernardes, Matthew Davies, and Carlos Guedes. *A Hierarchical Harmonic Mixing Method*, pages 151–170. 11 2018. Cited on pages 2 and 13.

[8] Gilberto Bernardes, Matthew E.P. Davies, and Carlos Guedes. A Pure Data Spectro-Morphological Analysis Toolkit for Sound-Based Composition. *Proceedings of 1st International Congress for Electroacoustic Music - Electroacoustic Winds 2015*, pages 31–38, 2015. Cited on page 12.

[9] D Bogdanov, Nicolas Wack, Emilia Gómez, Sankalp Gulati, P Herrera, O Mayor, G Roma, J Salamon, J Zapata, and Xavier Serra. ESSENTIA: An audio analysis library for music information retrieval. *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR 2013)*, (November):493–498, 2013. Cited on pages 9 and 37.

[10] William Brent. A Timbre Analysis And Classification Toolkit For Pure Data. *International Computer Music Conference, ICMC 2010*, 2010. Cited on page 26.

[11] John Brooke. SUS - A quick and dirty usability scale. 1986. Cited on page 48.

[12] Pedro Cano, Markus Koppenberger, and Nicolas Wack. Content-based music audio recommendation. *Proceedings of the 13th annual ACM international conference on Multimedia - MULTIMEDIA '05*, (May):211, 2005. Cited on page 10.

[13] Matteo Casu, Marinos Koutsomichalis, and Andrea Valle. Imaginary Soundscapes: The SoDA Project. *Proceedings of the 9th Audio Mostly: A Conference on Interaction With Sound*, 1:5:1—-5:8, 2014. Cited on pages 1 and 3.

[14] A. Correya. Retrieving Ambiguous Sounds Using Perceptual Timbral Attributes in Audio Production Environments, 2017. Cited on pages 2, 8, and 12.

[15] Marie Cottrell, Madalina Olteanu, Fabrice Rossi, and Nathalie Villa-vialaneix. Self-Organizing Maps , Theory and. 39(1):1–22, 2018. Cited on pages 13 and 14.

[16] Eduardo Coutinho and Angelo Cangelosi. Musical emotions: predicting second-by-second subjective feelings of emotion from low-level psychoacoustic features and physiological measurements. *Emotion*, 11(4):921, 2011. Cited on page 21.

[17] Kushal Dave and Vasudeva Varma. *Music Information Retrieval: Recent Developments and Applications*, volume 8. 2014. Cited on pages 2 and 10.

[18] Arne Eigenfeldt and Philippe Pasquier. Negotiated Content : Generative Soundscape Composition by Autonomous Musical Agents in Coming Together : Freesound Negotiated Content : Generative Soundscape Composition by Autonomous Musical Agents in Coming Together : Freesound Real-time Composition. (June 2014), 2011. Cited on page 16.

[19] Arne Eigenfeldt, Miles Thorogood, Jim Bizzocchi, and Philippe Pasquier. MediaScape: Towards a Video, Music, and Sound Metacreation. *Journal of Science and Technology of the Arts*, 6(1):61, 2014. Cited on page 8.

[20] Slim Essid and Thomas Fillon. YAAFE , an Easy to Use and Efficient Audio Feature Extraction Software . YAAFE , an easy to use and efficient audio feature extraction software. (January):7–13, 2010. Cited on pages 24 and 37.

[21] Brian S. Everitt, Sabine Landau, Morven Leese, and Daniel Stahl. *Cluster Analysis (5e)*. 2011. Cited on page 13.

[22] Jianyu Fan, Miles Thorogood, and Philippe Pasquier. Emo-Soundscapes: A Dataset for Soundscape Emotion Recognition. 2017. Cited on pages 21 and 22.

[23] Jianyu Fan, Miles Thorogood, Bernhard Riecke, and Philippe Pasquier. Automatic recognition of eventfulness and pleasantness of soundscape. 10 2015. Cited on pages 46 and 47.

[24] Nathaniel Finney. Autonomous generation of soundscapes using unstructured sound databases. *Communication*, 2009. Cited on page 8.

[25] Frederic Font and Giuseppe Bandiera. Freesound Explorer: Make Music While Discovering Freesound! *Proceedings of the WAC*, 2016. Cited on page 9.

[26] Jacek Grekow. Computer Information Systems and Industrial Management. 8838(September 2016), 2014. Cited on page 21.

[27] Thomas Grill. Perceptually Informed Organization of Textural Sounds. (June):123, 2012. Cited on page 12.

[28] Haotian Guan, Zhilei Liu, Longbiao Wang, Jianwu Dang, and Ruiguo Yu. Speech emotion recognition considering local dynamic features. In *International Seminar on Speech Production*, pages 14–23. Springer, 2017. Cited on page 21.

[29] P. Herrera and F. , Dehamel, A. , Gouyon. Automatic labeling of unpitched percussion sounds Perfecto. *New York*, pages 1–13, 2002. Cited on page 10.

[30] Perfecto Herrera, Xavier Serra, and Geoffroy Peeters. Audio descriptors and descriptor schemes in the context of MPEG-7. *International Computer Music Conference*, 1999. Cited on pages 1 and 10.

[31] Arefin Huq, Juan Pablo Bello, and Robert Rowe. Automated Music Emotion Recognition : A Systematic Evaluation. 39(3):227–244, 2010. Cited on page 22.

[32] Kandogan. Star Coordinates : A Multi-dimensional Visualization Technique with Uniform Treatment of Dimensions. *Star*, 650:1–4, 2000. Cited on page 13.

[33] VB Kobayashi and VB Calag. Detection of affective states from speech signals using ensembles of classifiers. 2013. Cited on page 21.

[34] P Kostagiolas. *Trends in Music Information Seeking, Behavior, and Retrieval for Creativity*. Advances in Multimedia and Interactive Technologies. IGI Global, 2016. Cited on page 1.

[35] Klaus Krippendorff. Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30(1):61–70, 1970. Cited on page 22.

[36] Peter J Lang, Margaret M Bradley, and Bruce N Cuthbert. Emotion, motivation, and anxiety: Brain mechanisms and psychophysiology. *Biological psychiatry*, 44(12):1248–1263, 1998. Cited on page 20.

[37] Randy J Larsen and Edward Diener. Promises and problems with the circumplex model of emotion. 1992. Cited on page 20.

[38] Olivier Lartillot, Petri Toiviainen, and Tuomas Eerola. A Matlab Toolbox for Music Information Retrieval A Matlab Toolbox for Music Information Retrieval. (June 2014), 2007. Cited on pages 24 and 37.

[39] Ming Zhe Li, Zhi Qing Hu, Zhong Yi Cai, and Xue Peng Gong. Method of efficient continuous plastic forming for freeform surface part. *Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)*, 37(3):489–494, 2007. Cited on page 22.

[40] Lu Lie, Zhang Hong-Jiang, and Jiang Hao. Content analysis for audio classification and segmentation. *IEEE Transactions on Speech and Audio Processing*, 10(7):504–516, 2002. Cited on page 10.

[41] Lev Manovich. What comes after remix? pages 1–6, 2007. Cited on page 1.

[42] Ananya Misra, Perry R Cook, and Ge Wang. Musical Tapestry : Re-composing Natural Sounds. 2002. Cited on page 15.

[43] Dalibor Mitrovic, Matthias Zeppelzauer, and Horst Eidenberger. Analysis of the data quality of audio descriptions of environmental sounds. *Journal of Digital Information Management*, 5(2):48, 2007. Cited on page 26.

[44] Cárthach Ó Nuanáin, Sergi Jordà, and Perfecto Herrera. An Interactive Software Instrument for Real-time Rhythmic Concatenative Synthesis. *New Interfaces for Musical Expression*, pages 383–387, 2016. Cited on pages 2 and 13.

[45] Hernán Ordiales and Matías Lennie Bruno. Sound recycling from public databases. *Proceedings of the 12th International Audio Mostly Conference on Augmented and Participatory Sound and Music Experiences - AM '17*, Part F1319(August):1–8, 2017. Cited on pages 8 and 16.

[46] J. Oswald. Bettered by the Borrower: The Ethics of Musical Debt, 2004. Cited on pages 1 and 2.

[47] Elias Pampalk, Andreas Rauber, and Dieter Merkl. Content-based organization and visualization of music archives. *Proceedings of the tenth ACM international conference on Multimedia - MULTIMEDIA '02*, page 570, 2002. Cited on pages 2 and 13.

[48] Renato Panda, Ricardo Manuel Malheiro, and Rui Pedro Paiva. Novel audio features for music emotion recognition. *IEEE Transactions on Affective Computing*, 2018. Cited on page 21.

[49] Conference Paper, Gilberto Bernardes, Carlos Guedes, and Bruce Pennycook. From Sounds to Music and Emotions. 7900(March 2016), 2013. Cited on pages 2, 12, 13, 14, and 15.

[50] Andy Pearce, Tim Brookes, and Russell Mason. Hierarchical ontology of timbral semantic descriptors. *AudioCommons - Deliverable D5.1*, pages 1–34, 2016. Cited on pages 2 and 12.

[51] Geoffroy Peeters. A large set of audio features for sound description (similarity and classification) in the cuidado project. 2004. Cited on page 24.

[52] Geoffroy Peeters, Bruno L. Giordano, Patrick Susini, Nicolas Misdariis, and Stephen McAdams. The Timbre Toolbox: Extracting audio descriptors from musical signals. *The Journal of the Acoustical Society of America*, 130(5):2902–2916, 2011. Cited on pages 12 and 26.

[53] A. Pierce. Audio Commons - Timbral explorer exploring sound effects with timbre. Cited on page 12.

[54] Ana Rodrigues, Gilberto Bernardes, Penousal Machado, and Catarina Mac. MixMash : A Visualisation System for Musical Mashup Creation. (July), 2018. Cited on pages 2 and 13.

[55] James A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980. Cited on pages xi, 19, 20, 22, and 29.

[56] James A Russell and James M Carroll. On the bipolarity of positive and negative affect. *Psychological bulletin*, 125(1):3, 1999. Cited on pages 19 and 20.

[57] James A Russell and Beverly Fehr. Fuzzy concepts in a fuzzy hierarchy: Varieties of anger. *Journal of personality and social psychology*, 67(2):186, 1994. Cited on page 19.

[58] Carolyn Saarni. A skill-based model of emotional competence: A developmental perspective. 1999. Cited on page 19.

[59] P Schaeffer. *Treatise on Musical Objects: An Essay across Disciplines*. California Studies in 20th-Century Music. University of California Press, 2017. Cited on pages 1 and 12.

[60] R M Schafer. *The new soundscape: a handbook for the modern music teacher*. BMI Canada, 1969. Cited on page 7.

[61] R M Schafer. *The Book of Noise*. Collection Serge-Garant: Monographies. Price Milburn, 1970. Cited on page 7.

[62] R M Schafer. *The Tuning of the World*. Borzoi book. Knopf, 1977. Cited on pages xi, 3, 7, 8, 21, 41, and 47.

[63] Emery Schubert and Joe Wolfe. Does timbral brightness scale with frequency and spectral centroid. *Acustica*, 92:820–, 09 2006. Cited on page 12.

[64] Diemo Schwarz. A System for Data-Driven Concatenative Sound Synthesis. *Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-00)*, pages 97–102, 2000. Cited on page 10.

[65] Diemo Schwarz, Grégory Beller, Bruno Verbrugghe, and Sam Britton. Real-time Corpus-Based Concatenative Synthesis with CataRT. *Proceedings of the 9th Int. Conference on Digital Audio Effects (DAFx-06), Montreal, Canada*, (September 2006):1–7, 2006. Cited on pages 2 and 12.

[66] Yading Song, Simon Dixon, and Marcus Pearce. Evaluation of musical features for emotion classification. *13th International Society for Music Information Retrieval Conference*, (Ismir):523–528, 2012. Cited on page 14.

[67] Kıvanç Tatar and Philippe Pasquier. Musical agents: A typology and state of the art towards musical metacreation. *Journal of New Music Research*, 48(1):56–105, 2019. Cited on page 56.

[68] Robert E Thayer. *The biopsychology of mood and arousal*. Oxford University Press, 1990. Cited on page 20.

[69] Miles Thorogood, Jianyu Fan, and Philippe Pasquier. Soundscape audio signal classification and segmentation using listeners perception of background and foreground sound. *Journal of the Audio Engineering Society*, 64(7/8):484–492, 2016. Cited on page 21.

[70] Miles Thorogood, Philippe Pasquier, and Arne Eigenfeldt. Audio Metaphor: Audio Information Retrieval for Soundscape Composition. *Smc 2012*, (March 2016):372–378, 2012. Cited on page 8.

[71] Luca Turchet and Mathieu Barthet. Jamming with a Smart Mandolin and Freesound-based Accompaniment. *Conference of Open Innovation Association, FRUCT*, 2018-November(November):375–381, 2018. Cited on pages 1 and 8.

[72] Andrea Valle, Matteo Casu, and Paolo Armao. SoDA : A Sound Design Accelerator for the automatic generation of soundscapes from an ontologically annotated sound library. *Proceedings ICMC|SMC|2014*, (September):14–20, 2014. Cited on page 3.

[73] Barry Vercoe. MIT Media Lab Perceptual Sound Synthesizer A Study of Auditory Perception Through Language Ananth Ram. 2006. Cited on page 12.

[74] Hugues Vinet, Perfecto Herrera, and François Pachet. The CUIDADO Project. *3rd International Society for Music Information Retrieval (ISMIR) Conference*, (May 2014):197–203, 2002. Cited on pages 10 and 11.

[75] David Watson and Lee Anna Clark. Affects separable and inseparable: on the hierarchical arrangement of the negative affects. *Journal of personality and social psychology*, 62(3):489, 1992. Cited on page 19.

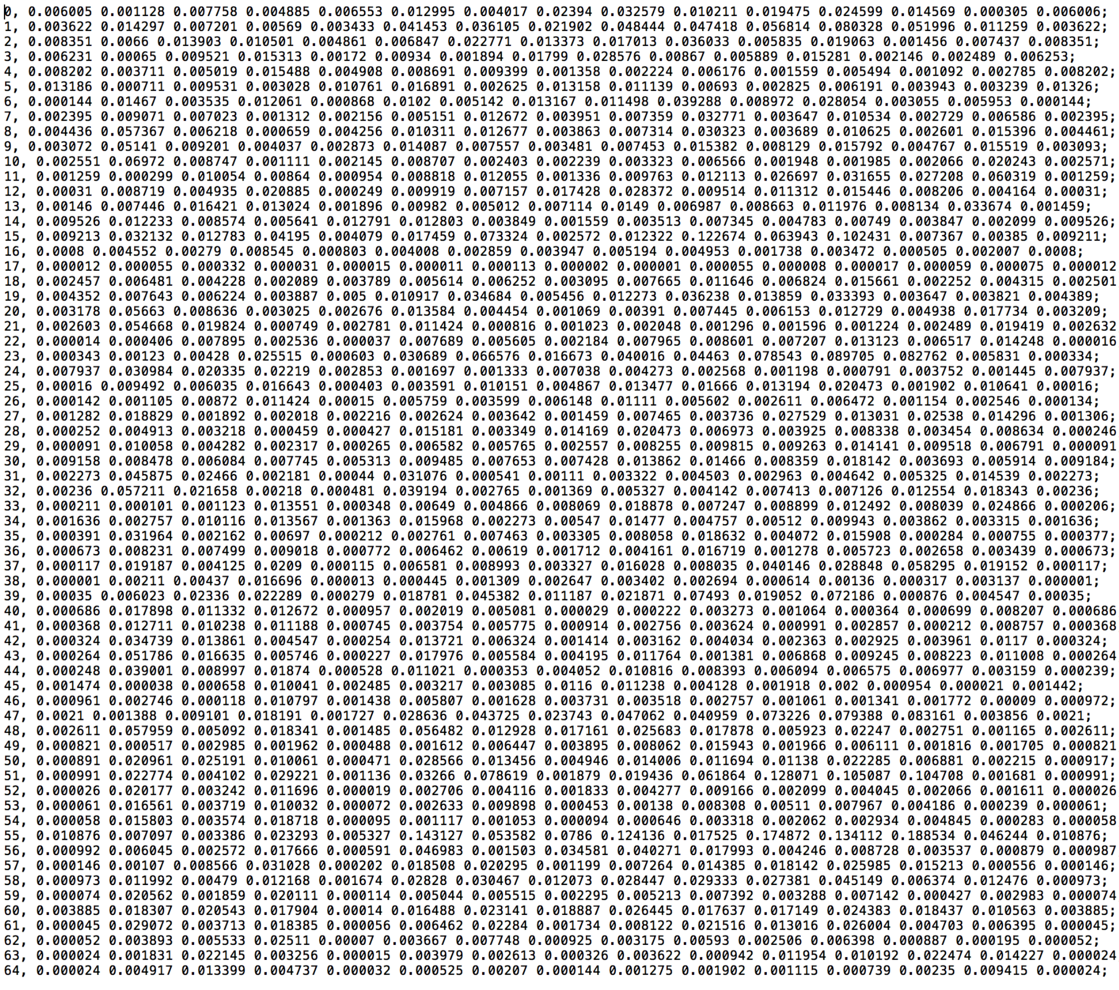
[76] David Watson, David Wiese, Jatin Vaidya, and Auke Tellegen. The two general activation systems of affect: Structural findings, evolutionary considerations, and psychobiological evidence. *Journal of personality and social psychology*, 76(5):820, 1999. Cited on page 20.

[77] Anna Xambó, Johan Pauwels, Gerard Roma, Mathieu Barthet, György Fazekas, and Carlos A Ballesteros. Jam with Jamendo : Querying a Large Music Collection by Chords from a Learner ' s Perspective. (Cc):1–7, 2018. Cited on page 8.

Appendix A

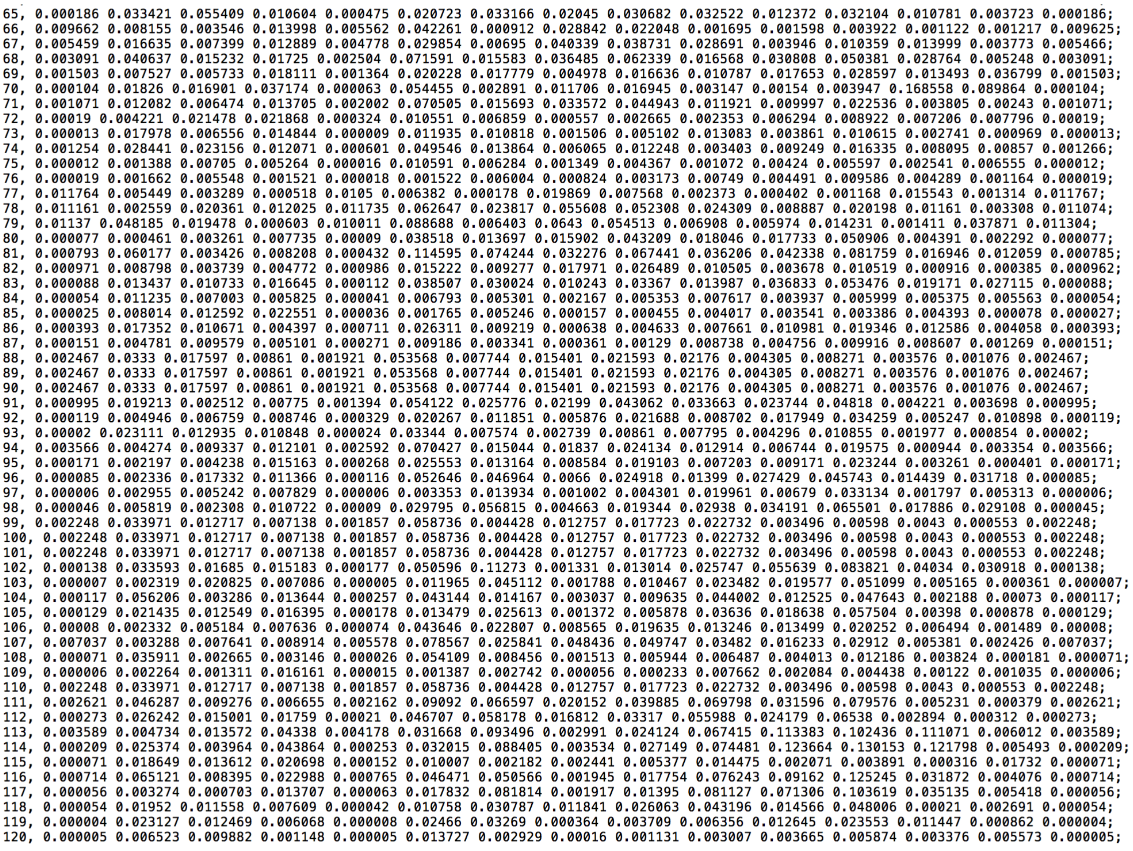
Appendix

A.1 Descriptors Relevancy



```
0, 0.006005 0.001128 0.007758 0.004885 0.006553 0.012995 0.004017 0.02394 0.032579 0.010211 0.019475 0.024599 0.014569 0.000305 0.006006;
1, 0.003622 0.014297 0.007201 0.00569 0.003433 0.041453 0.036105 0.021902 0.048444 0.047418 0.056814 0.080328 0.051996 0.011259 0.003622;
2, 0.008351 0.0066 0.013903 0.010501 0.004861 0.006847 0.022771 0.013373 0.017013 0.036033 0.005835 0.019063 0.001456 0.007437 0.008351;
3, 0.006231 0.00065 0.009521 0.015313 0.00172 0.00934 0.001894 0.01799 0.028576 0.00867 0.005889 0.015281 0.002146 0.002489 0.006253;
4, 0.008202 0.003711 0.005019 0.015488 0.004908 0.008691 0.009399 0.001358 0.002224 0.006176 0.001559 0.005494 0.001092 0.002785 0.008202;
5, 0.013186 0.000711 0.009531 0.003028 0.010761 0.016891 0.002625 0.013158 0.011139 0.00693 0.002825 0.006191 0.003943 0.003239 0.01326;
6, 0.000144 0.01467 0.003535 0.012061 0.000868 0.0102 0.005142 0.013167 0.011498 0.039288 0.008972 0.028054 0.003055 0.005953 0.000144;
7, 0.002395 0.009071 0.007023 0.001312 0.002156 0.005151 0.012672 0.003951 0.007359 0.032771 0.003647 0.010534 0.002729 0.006586 0.002395;
8, 0.004436 0.057367 0.006218 0.000659 0.004256 0.010311 0.012677 0.003863 0.007314 0.030323 0.003689 0.010625 0.002601 0.015396 0.004461;
9, 0.003072 0.05141 0.009201 0.004037 0.002873 0.014087 0.007557 0.003481 0.007453 0.015382 0.008129 0.015792 0.004767 0.015519 0.003093;
10, 0.002551 0.06972 0.008747 0.001111 0.002145 0.008707 0.002403 0.002239 0.003323 0.006566 0.001948 0.001985 0.002066 0.020243 0.002571;
11, 0.001259 0.000299 0.010054 0.00864 0.000954 0.008818 0.012055 0.001336 0.009763 0.012113 0.026697 0.031655 0.027208 0.060319 0.001259;
12, 0.00031 0.008719 0.004935 0.020885 0.000249 0.009919 0.007157 0.017428 0.028372 0.009514 0.011312 0.015446 0.008206 0.004164 0.00031;
13, 0.00146 0.007446 0.016421 0.013024 0.001896 0.00982 0.005012 0.007114 0.0149 0.006987 0.008663 0.011976 0.008134 0.033674 0.001459;
14, 0.009526 0.012233 0.008574 0.005641 0.012791 0.012803 0.003849 0.001559 0.003513 0.007345 0.004783 0.00749 0.003847 0.002099 0.009526;
15, 0.009213 0.032132 0.012783 0.04195 0.004079 0.017459 0.073324 0.002572 0.012322 0.122674 0.063943 0.102431 0.007367 0.00385 0.009211;
16, 0.0008 0.004552 0.00279 0.008545 0.000803 0.004008 0.002859 0.003947 0.005194 0.004953 0.001738 0.003472 0.000505 0.002007 0.0008;
17, 0.000012 0.000055 0.000332 0.000031 0.000015 0.000011 0.000113 0.000002 0.000001 0.000055 0.000008 0.000017 0.000059 0.000075 0.000012;
18, 0.002457 0.006481 0.004228 0.002089 0.003789 0.005614 0.006252 0.003095 0.007665 0.011646 0.006824 0.015661 0.002252 0.004315 0.002501;
19, 0.004352 0.007643 0.006224 0.003887 0.005 0.010917 0.034684 0.005456 0.012273 0.036238 0.013859 0.033393 0.003647 0.003821 0.004389;
20, 0.003178 0.05663 0.008636 0.003025 0.002676 0.013584 0.004454 0.001069 0.00391 0.007445 0.006153 0.012729 0.004938 0.017734 0.003209;
21, 0.002603 0.054668 0.019824 0.000749 0.002781 0.011424 0.000816 0.001023 0.002048 0.001296 0.001596 0.001224 0.002489 0.019419 0.002632;
22, 0.000014 0.000406 0.007895 0.002536 0.000037 0.007689 0.005605 0.002184 0.007965 0.008601 0.007207 0.013123 0.006517 0.014248 0.000016;
23, 0.000343 0.00123 0.00428 0.025515 0.000603 0.030689 0.066576 0.016673 0.040016 0.04463 0.078543 0.089705 0.082762 0.005831 0.000334;
24, 0.007937 0.030984 0.020335 0.02219 0.002853 0.001697 0.001333 0.007038 0.004273 0.002568 0.001198 0.000791 0.003752 0.001445 0.007937;
25, 0.00016 0.009492 0.006035 0.016643 0.000403 0.003591 0.010151 0.004867 0.013477 0.01666 0.013194 0.020473 0.001902 0.010641 0.00016;
26, 0.000142 0.001105 0.00872 0.011424 0.00015 0.005759 0.003599 0.006148 0.01111 0.005602 0.002611 0.006472 0.001154 0.002546 0.000134;
27, 0.001282 0.018829 0.001892 0.002018 0.002216 0.002624 0.003642 0.001459 0.007465 0.003736 0.027529 0.013031 0.02538 0.014296 0.001306;
28, 0.000252 0.004913 0.003218 0.000459 0.000427 0.015181 0.003349 0.014169 0.020473 0.006973 0.003925 0.008338 0.003454 0.008634 0.000246;
29, 0.000091 0.010058 0.004282 0.002317 0.000265 0.006582 0.005765 0.002557 0.008255 0.009815 0.009263 0.014141 0.009518 0.006791 0.000091;
30, 0.009158 0.008478 0.006084 0.007745 0.005313 0.009485 0.007653 0.007428 0.013862 0.01466 0.008359 0.018142 0.003693 0.005914 0.009184;
31, 0.002273 0.045875 0.02466 0.002181 0.00044 0.031076 0.000541 0.00111 0.003322 0.004503 0.002963 0.004642 0.005325 0.014539 0.002273;
32, 0.00236 0.057211 0.021658 0.00218 0.000481 0.039194 0.002765 0.001369 0.005327 0.004142 0.007413 0.007126 0.012554 0.018343 0.00236;
33, 0.000211 0.000101 0.001123 0.013551 0.000348 0.00649 0.004866 0.008069 0.018878 0.007247 0.008899 0.012492 0.008039 0.024866 0.000206;
34, 0.001636 0.002757 0.010116 0.013567 0.001363 0.015968 0.002273 0.00547 0.01477 0.004757 0.00512 0.009943 0.003862 0.003315 0.001636;
35, 0.000391 0.031964 0.002162 0.00697 0.000212 0.002761 0.007463 0.003305 0.008058 0.018632 0.004072 0.015908 0.000284 0.000755 0.000377;
36, 0.000673 0.008231 0.007499 0.009018 0.000772 0.006462 0.00619 0.001712 0.004161 0.016719 0.001278 0.005723 0.002658 0.003439 0.000673;
37, 0.000117 0.019187 0.004125 0.0209 0.000115 0.006581 0.008993 0.003327 0.016028 0.008035 0.040146 0.028848 0.058295 0.019152 0.000117;
38, 0.000001 0.00211 0.00437 0.016696 0.000013 0.000445 0.001309 0.002647 0.003402 0.002694 0.000614 0.00136 0.000317 0.003137 0.000001;
39, 0.00035 0.006023 0.02336 0.022289 0.000279 0.018781 0.045382 0.011187 0.021871 0.07493 0.019052 0.072186 0.000876 0.004547 0.00035;
40, 0.000686 0.017898 0.011332 0.012672 0.000957 0.002019 0.005081 0.000029 0.000222 0.003273 0.001064 0.000364 0.000699 0.008207 0.000686;
41, 0.000368 0.012711 0.010238 0.011188 0.000745 0.003754 0.005775 0.000914 0.002756 0.003624 0.000991 0.002857 0.000212 0.008757 0.000368;
42, 0.000324 0.034739 0.013861 0.004547 0.000254 0.013721 0.006324 0.001414 0.003162 0.004034 0.002363 0.002925 0.003961 0.0117 0.000324;
43, 0.000264 0.051786 0.016635 0.005746 0.000227 0.017976 0.005584 0.004195 0.011764 0.001381 0.006868 0.009245 0.008223 0.011008 0.000264;
44, 0.000248 0.039001 0.008997 0.01874 0.000528 0.011021 0.000353 0.004052 0.010816 0.008393 0.006094 0.006575 0.006977 0.003159 0.000239;
45, 0.001474 0.000038 0.000658 0.010041 0.002485 0.003217 0.003085 0.0116 0.011238 0.004128 0.001918 0.002 0.000954 0.000021 0.001442;
46, 0.000961 0.002746 0.000118 0.010797 0.001438 0.005807 0.001628 0.003731 0.003518 0.002757 0.001061 0.001341 0.001772 0.00009 0.000972;
47, 0.0021 0.001388 0.009101 0.018191 0.001727 0.028636 0.043725 0.023743 0.047062 0.040959 0.073226 0.079388 0.083161 0.003856 0.0021;
48, 0.002611 0.057959 0.005092 0.018341 0.001485 0.056482 0.012928 0.017161 0.025683 0.017878 0.005923 0.02247 0.002751 0.001165 0.002611;
49, 0.000821 0.000517 0.002985 0.001962 0.000488 0.001612 0.006447 0.003895 0.008062 0.015943 0.001966 0.006111 0.001816 0.001705 0.000821;
50, 0.000891 0.020961 0.025191 0.010061 0.000471 0.028566 0.013456 0.004946 0.014006 0.011694 0.01138 0.022285 0.006881 0.002215 0.000917;
51, 0.000991 0.022774 0.004102 0.029221 0.001136 0.03266 0.078619 0.001879 0.019436 0.061864 0.128071 0.105087 0.104708 0.001681 0.000991;
52, 0.000026 0.020177 0.003242 0.011696 0.000019 0.002706 0.004116 0.001833 0.004277 0.009166 0.002099 0.004045 0.002066 0.001611 0.000026;
53, 0.000061 0.016561 0.003719 0.010032 0.000072 0.002633 0.009898 0.000453 0.00138 0.008308 0.00511 0.007967 0.004186 0.000239 0.000061;
54, 0.000058 0.015803 0.003574 0.018718 0.000095 0.001117 0.001053 0.000094 0.000646 0.003318 0.002062 0.002934 0.004845 0.000283 0.000058;
55, 0.010876 0.007097 0.003386 0.023293 0.005327 0.143127 0.053582 0.0786 0.124136 0.017525 0.174872 0.134112 0.188534 0.046244 0.010876;
56, 0.000992 0.006045 0.002572 0.017666 0.000591 0.046983 0.001503 0.034581 0.040271 0.017993 0.004246 0.008728 0.003537 0.000879 0.000987;
57, 0.000146 0.00107 0.008566 0.031028 0.000202 0.018508 0.020295 0.001199 0.007264 0.014385 0.018142 0.025985 0.015213 0.000556 0.000146;
58, 0.000973 0.011992 0.00479 0.012168 0.001674 0.02828 0.030467 0.012073 0.028447 0.029333 0.027381 0.045149 0.006374 0.012476 0.000973;
59, 0.000074 0.020562 0.001859 0.020111 0.000114 0.005044 0.005515 0.002295 0.005213 0.007392 0.003288 0.007142 0.000427 0.002983 0.000074;
60, 0.003885 0.018307 0.020543 0.017904 0.00014 0.016488 0.023141 0.018887 0.026445 0.017637 0.017149 0.024383 0.018437 0.010563 0.003885;
61, 0.000045 0.029072 0.003713 0.018385 0.000056 0.006462 0.02284 0.001734 0.008122 0.021516 0.013016 0.026004 0.004703 0.006395 0.000045;
62, 0.000052 0.003893 0.005533 0.02511 0.00007 0.003667 0.007748 0.000925 0.003175 0.00593 0.002506 0.006398 0.000887 0.000195 0.000052;
63, 0.000024 0.001831 0.022145 0.003256 0.000015 0.003979 0.002613 0.000326 0.003622 0.000942 0.011954 0.010192 0.022474 0.014227 0.000024;
64, 0.000024 0.004917 0.013399 0.004737 0.000032 0.000525 0.00207 0.000144 0.001275 0.001902 0.001115 0.000739 0.00235 0.009415 0.000024;
```

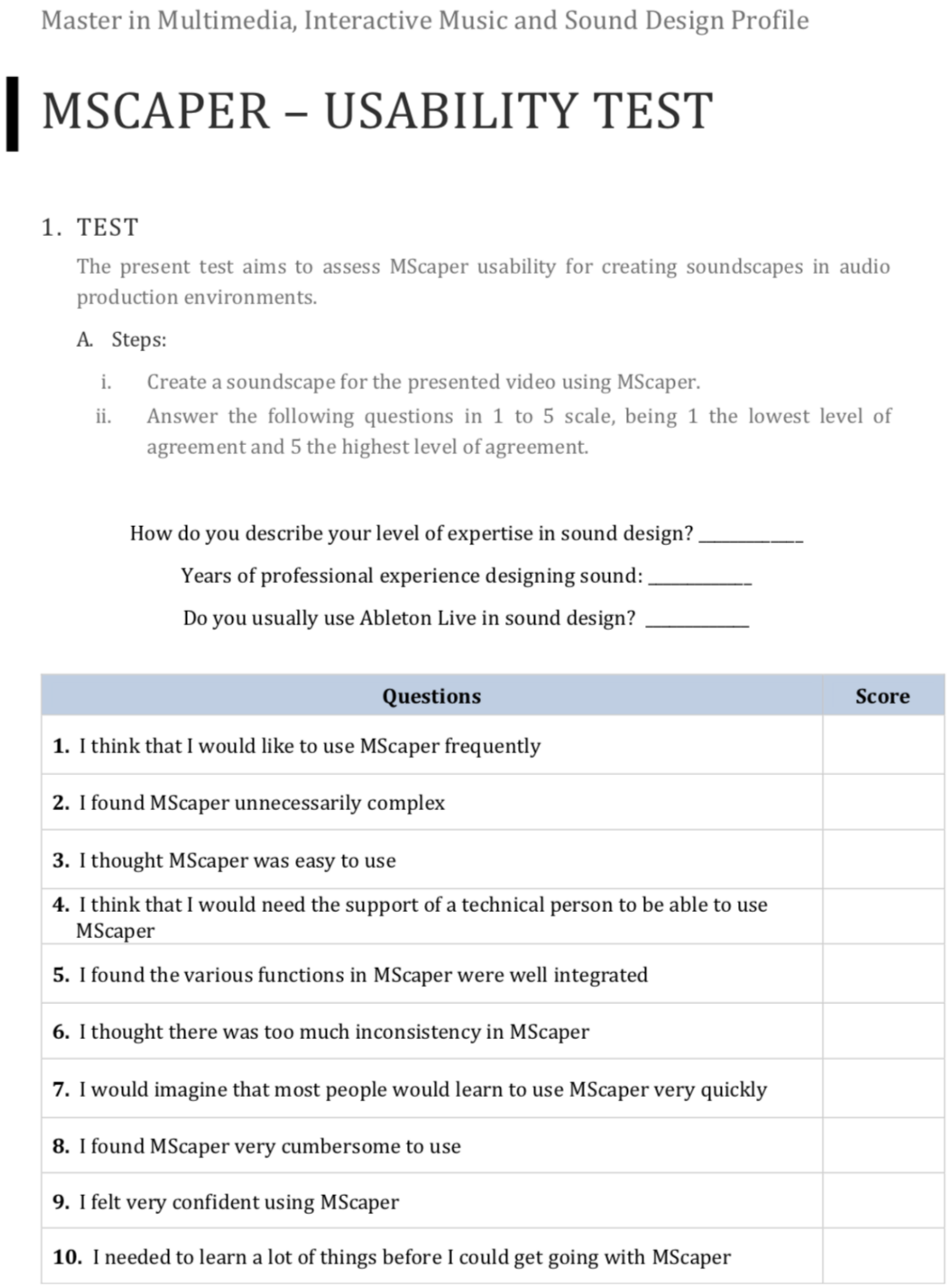
Figure A.1: Mean Descriptors Variance (Relevancy) of the Quantized Space. Part 1



```
65, 0.000186 0.033421 0.055409 0.010604 0.000475 0.020723 0.033166 0.02045 0.030682 0.032522 0.012372 0.032104 0.010781 0.003723 0.000186;
66, 0.009662 0.008155 0.003546 0.013998 0.005562 0.042261 0.000912 0.028842 0.022048 0.001695 0.001598 0.003922 0.001122 0.001217 0.009625;
67, 0.005459 0.016635 0.007399 0.012889 0.004778 0.029854 0.00695 0.040339 0.038731 0.028691 0.003946 0.010359 0.013999 0.003773 0.005466;
68, 0.003091 0.040637 0.015232 0.01725 0.002504 0.071591 0.015583 0.036485 0.062339 0.016568 0.030808 0.050381 0.028764 0.005248 0.003091;
69, 0.001503 0.007527 0.005733 0.018111 0.001364 0.020228 0.017779 0.004978 0.016636 0.010787 0.017653 0.028597 0.013493 0.036799 0.001503;
70, 0.000104 0.01826 0.016901 0.037174 0.000063 0.054455 0.002891 0.011706 0.016945 0.003147 0.00154 0.003947 0.168558 0.089864 0.000104;
71, 0.001071 0.012082 0.006474 0.013705 0.002002 0.070505 0.015693 0.033572 0.044943 0.011921 0.009997 0.022536 0.003805 0.00243 0.001071;
72, 0.00019 0.004221 0.021478 0.021868 0.000324 0.010551 0.006859 0.000557 0.002665 0.002353 0.006294 0.008922 0.007206 0.007796 0.00019;
73, 0.000013 0.017978 0.006556 0.014844 0.000009 0.011935 0.010818 0.001506 0.005102 0.013083 0.003861 0.010615 0.002741 0.000969 0.000013;
74, 0.001254 0.028441 0.023156 0.012071 0.000601 0.049546 0.013864 0.006065 0.012248 0.003403 0.009249 0.016335 0.008095 0.00857 0.001266;
75, 0.000012 0.001388 0.00705 0.005264 0.000016 0.010591 0.006284 0.001349 0.004367 0.001072 0.00424 0.005597 0.002541 0.006555 0.000012;
76, 0.000019 0.001662 0.005548 0.001521 0.000018 0.001522 0.006004 0.000824 0.003173 0.00749 0.004491 0.009586 0.004289 0.001164 0.000019;
77, 0.011764 0.005449 0.003289 0.000518 0.0105 0.006382 0.000178 0.019869 0.007568 0.002373 0.000402 0.001168 0.015543 0.001314 0.011767;
78, 0.011161 0.002559 0.020361 0.012025 0.011735 0.062647 0.023817 0.055608 0.052308 0.024309 0.008887 0.020198 0.01161 0.003308 0.011074;
79, 0.01137 0.048185 0.019478 0.000603 0.010011 0.088688 0.006403 0.0643 0.054513 0.006908 0.005974 0.014231 0.001411 0.037871 0.011304;
80, 0.000077 0.000461 0.003261 0.007735 0.00009 0.038518 0.013697 0.015902 0.043209 0.018046 0.017733 0.050906 0.004391 0.002292 0.000077;
81, 0.000793 0.060177 0.003426 0.008208 0.000432 0.114595 0.074244 0.032276 0.067441 0.036206 0.042338 0.081759 0.016946 0.012059 0.000785;
82, 0.000971 0.008798 0.003739 0.004772 0.000986 0.015222 0.009277 0.017971 0.026489 0.010505 0.003678 0.010519 0.000916 0.000385 0.000962;
83, 0.000088 0.013437 0.010733 0.016645 0.000112 0.038507 0.030024 0.010243 0.03367 0.013987 0.036833 0.053476 0.019171 0.027115 0.000088;
84, 0.000054 0.011235 0.007003 0.005825 0.000041 0.006793 0.005301 0.002167 0.005353 0.007617 0.003937 0.005999 0.005375 0.005563 0.000054;
85, 0.000025 0.008014 0.012592 0.022551 0.000036 0.001765 0.005246 0.000157 0.000455 0.004017 0.003541 0.003386 0.004393 0.000078 0.000027;
86, 0.000393 0.017352 0.010671 0.004397 0.000711 0.026311 0.009219 0.000638 0.004633 0.007661 0.010981 0.019346 0.012586 0.004058 0.000393;
87, 0.000151 0.004781 0.009579 0.005101 0.000271 0.009186 0.003341 0.000361 0.00129 0.008738 0.004756 0.009916 0.008607 0.001269 0.000151;
88, 0.002467 0.0333 0.017597 0.00861 0.001921 0.053568 0.007744 0.015401 0.021593 0.02176 0.004305 0.008271 0.003576 0.001076 0.002467;
89, 0.002467 0.0333 0.017597 0.00861 0.001921 0.053568 0.007744 0.015401 0.021593 0.02176 0.004305 0.008271 0.003576 0.001076 0.002467;
90, 0.002467 0.0333 0.017597 0.00861 0.001921 0.053568 0.007744 0.015401 0.021593 0.02176 0.004305 0.008271 0.003576 0.001076 0.002467;
91, 0.000995 0.019213 0.002512 0.00775 0.001394 0.054122 0.025776 0.02199 0.043062 0.033663 0.023744 0.04818 0.004221 0.003698 0.000995;
92, 0.000119 0.004946 0.006759 0.008746 0.000329 0.020267 0.011851 0.005876 0.021688 0.008702 0.017949 0.034259 0.005247 0.010898 0.000119;
93, 0.00002 0.023111 0.012935 0.010848 0.000024 0.03344 0.007574 0.002739 0.00861 0.007795 0.004296 0.010855 0.001977 0.000854 0.00002;
94, 0.003566 0.004274 0.009337 0.012101 0.002592 0.070427 0.015044 0.01837 0.024134 0.012914 0.006744 0.019575 0.000944 0.003354 0.003566;
95, 0.000171 0.002197 0.004238 0.015163 0.000268 0.025553 0.013164 0.008584 0.019103 0.007203 0.009171 0.023244 0.003261 0.000401 0.000171;
96, 0.000085 0.002336 0.017332 0.011366 0.000116 0.052646 0.046964 0.0066 0.024918 0.01399 0.027429 0.045743 0.014439 0.031718 0.000085;
97, 0.000006 0.002955 0.005242 0.007829 0.000006 0.003353 0.013934 0.001002 0.004301 0.019961 0.00679 0.033134 0.001797 0.005313 0.000006;
98, 0.000046 0.005819 0.002308 0.010722 0.00009 0.029795 0.056815 0.004663 0.019344 0.02938 0.034191 0.065501 0.017886 0.029108 0.000045;
99, 0.002248 0.033971 0.012717 0.007138 0.001857 0.058736 0.004428 0.012757 0.017723 0.022732 0.003496 0.00598 0.0043 0.000553 0.002248;
100, 0.002248 0.033971 0.012717 0.007138 0.001857 0.058736 0.004428 0.012757 0.017723 0.022732 0.003496 0.00598 0.0043 0.000553 0.002248;
101, 0.002248 0.033971 0.012717 0.007138 0.001857 0.058736 0.004428 0.012757 0.017723 0.022732 0.003496 0.00598 0.0043 0.000553 0.002248;
102, 0.000138 0.033593 0.01685 0.015183 0.000177 0.050596 0.11273 0.001331 0.013014 0.025747 0.055639 0.083821 0.04034 0.030918 0.000138;
103, 0.000007 0.002319 0.020825 0.007086 0.000005 0.011965 0.045112 0.001788 0.010467 0.023482 0.019577 0.051099 0.005165 0.000361 0.000007;
104, 0.000117 0.056206 0.003286 0.013644 0.000257 0.043144 0.014167 0.003037 0.009635 0.044002 0.012525 0.047643 0.002188 0.00073 0.000117;
105, 0.000129 0.021435 0.012549 0.016395 0.000178 0.013479 0.025613 0.001372 0.005878 0.03636 0.018638 0.057504 0.00398 0.000878 0.000129;
106, 0.00008 0.002332 0.005184 0.007636 0.000074 0.043646 0.022807 0.008565 0.019635 0.013246 0.013499 0.020252 0.006494 0.001489 0.00008;
107, 0.007037 0.003288 0.007641 0.008914 0.005578 0.078567 0.025841 0.048436 0.049747 0.03482 0.016233 0.02912 0.005381 0.002426 0.007037;
108, 0.000071 0.035911 0.002665 0.003146 0.000026 0.054109 0.008456 0.001513 0.005944 0.006487 0.004013 0.012186 0.003824 0.000181 0.000071;
109, 0.000006 0.002264 0.001311 0.016161 0.000015 0.001387 0.002742 0.000056 0.000233 0.007662 0.002084 0.004438 0.00122 0.001035 0.000006;
110, 0.002248 0.033971 0.012717 0.007138 0.001857 0.058736 0.004428 0.012757 0.017723 0.022732 0.003496 0.00598 0.0043 0.000553 0.002248;
111, 0.002621 0.046287 0.009276 0.006655 0.002162 0.09092 0.066597 0.020152 0.039885 0.069798 0.031596 0.079576 0.005231 0.000379 0.002621;
112, 0.000273 0.026242 0.015001 0.01759 0.00021 0.046707 0.058178 0.016812 0.03317 0.055988 0.024179 0.06538 0.002894 0.000312 0.000273;
113, 0.003589 0.004734 0.013572 0.04338 0.004178 0.031668 0.093496 0.002991 0.024124 0.067415 0.113383 0.102436 0.111071 0.006012 0.003589;
114, 0.000209 0.025374 0.003964 0.043864 0.000253 0.032015 0.088405 0.003534 0.027149 0.074481 0.123664 0.130153 0.121798 0.005493 0.000209;
115, 0.000071 0.018649 0.013612 0.020698 0.000152 0.010007 0.002182 0.002441 0.005377 0.014475 0.002071 0.003891 0.000316 0.01732 0.000071;
116, 0.000714 0.065121 0.008395 0.022988 0.000765 0.046471 0.050566 0.001945 0.017754 0.076243 0.09162 0.125245 0.031872 0.004076 0.000714;
117, 0.000056 0.003274 0.000703 0.013707 0.000063 0.017832 0.081814 0.001917 0.01395 0.081127 0.071306 0.103619 0.035135 0.005418 0.000056;
118, 0.000054 0.01952 0.011558 0.007609 0.000042 0.010758 0.030787 0.011841 0.026063 0.043196 0.014566 0.048006 0.00021 0.002691 0.000054;
119, 0.000004 0.023127 0.012469 0.006068 0.000008 0.02466 0.03269 0.000364 0.003709 0.006356 0.012645 0.023553 0.011447 0.000862 0.000004;
120, 0.000005 0.006523 0.009882 0.001148 0.000005 0.013727 0.002929 0.00016 0.001131 0.003007 0.003665 0.005874 0.003376 0.005573 0.000005;
```

Figure A.2: Mean Descriptors Variance (Relevancy) of the Quantized Space. Part 2

A.2 System Usability Scale



Master in Multimedia, Interactive Music and Sound Design Profile

MSCAPER – USABILITY TEST

1. TEST

The present test aims to assess MScaper usability for creating soundscapes in audio production environments.

A. Steps:

i. Create a soundscape for the presented video using MScaper.

ii. Answer the following questions in 1 to 5 scale, being 1 the lowest level of agreement and 5 the highest level of agreement.

How do you describe your level of expertise in sound design? ___________

Years of professional experience designing sound: ___________

Do you usually use Ableton Live in sound design? ___________

Questions	Score
1. I think that I would like to use MScaper frequently	
2. I found MScaper unnecessarily complex	
3. I thought MScaper was easy to use	
4. I think that I would need the support of a technical person to be able to use MScaper	
5. I found the various functions in MScaper were well integrated	
6. I thought there was too much inconsistency in MScaper	
7. I would imagine that most people would learn to use MScaper very quickly	
8. I found MScaper very cumbersome to use	
9. I felt very confident using MScaper	
10. I needed to learn a lot of things before I could get going with MScaper	

Figure A.3: SUS Test Answer Sheet

A.3 Perceptual Test: Descriptive Statistics

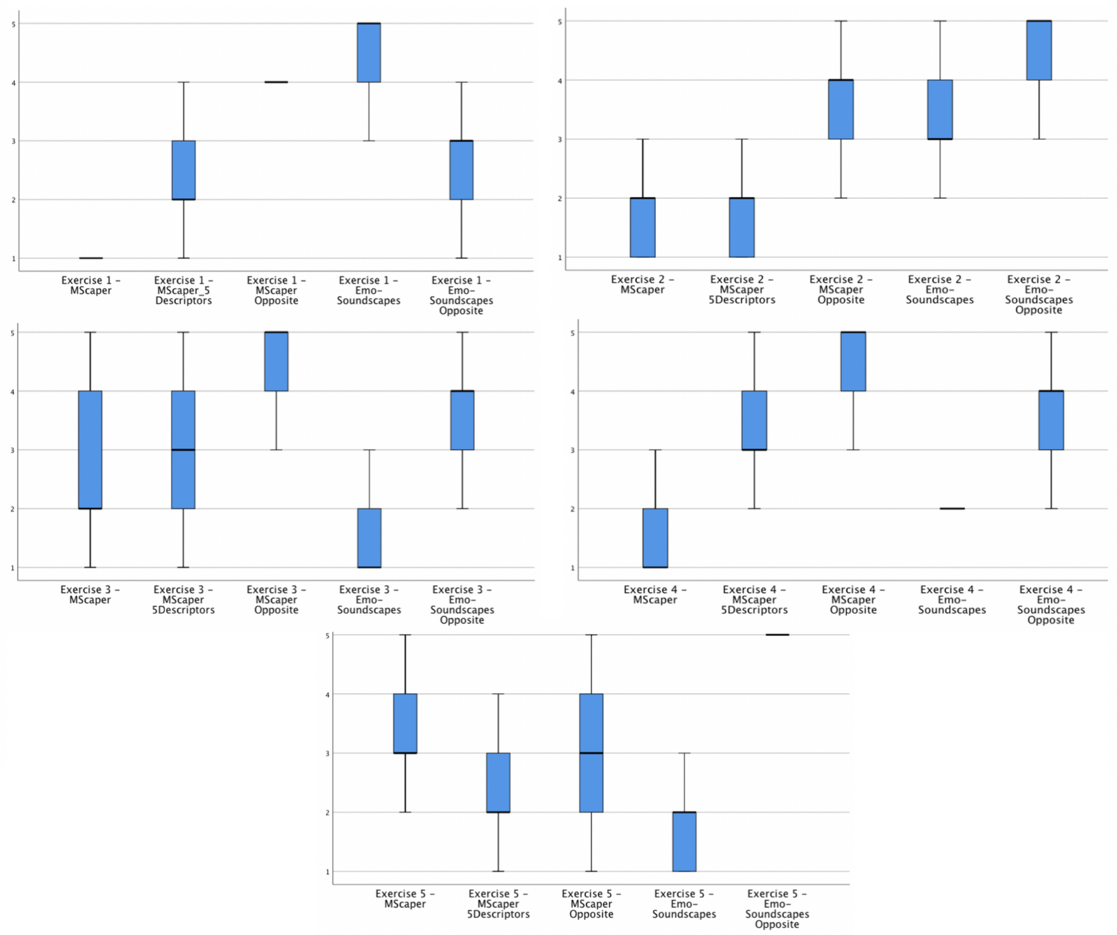


Figure A.4: Descriptive Statistics from All Questions of Perceptual Test