



TECHNISCHE UNIVERSITÄT BERLIN

Fakultät I - Geisteswissenschaften  
Fachgebiet Audiokommunikation  
Audiokommunikation und -technologie M.Sc.

# SELF-ORGANIZING MAPS FOR SOUND CORPUS ORGANIZATION

Master's Thesis

<b>Vorgelegt von:</b>	Jonas Margraf
<b>Erstgutachter:</b>	Prof. Dr. Stefan Weinzierl
<b>Zweitgutachter:</b>	Dr. Diemo Schwarz
<b>Datum:</b>	1. April 2019



## Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und eigenhändig sowie ohne unerlaubte fremde Hilfe und ausschließlich unter Verwendung der aufgeführten Quellen und Hilfsmittel angefertigt habe.

Berlin, den 1. April 2019

.....  
Jonas Margraf

## Abstract

Large collections of audio files — *sound corpora* — have never been more readily available. Sample libraries are easily accessible online and cheap storage media effectively eradicate concerns of storage capacity for contemporary music producers. At the same time, tools for navigating, searching and organizing these increasingly unmanageable audio file collections have not kept pace. At present, arguably the most common tool with which producers search their sample libraries are file browsers that simply present lists of file names in alphabetical order.

The present thesis approaches this problem from a practical perspective. We implement the Self-Organizing Map (SOM), an established machine learning algorithm for dimensionality reduction and data visualization, and apply it to sound corpus organization. We present *SOM Browser*, a fast, visual interface for sample library exploration. It offers an alternative to the established music production workflow by incorporating the SOM algorithm, which to our knowledge is not available in any commercial audio software. It is a standalone application that organizes a collection of sound files completely unsupervised and presents a two-dimensional map of the sounds. The map forms an interactive grid interface with which the user can audition files in rapid succession. This allows for a quick way to gain an overview of the analyzed sound corpus. To optimize the space allotted to the map interface, we extend the SOM algorithm with a new method, which we call Forced Node Population (FNP). FNP reduces unpopulated (“empty”) areas of the map at the cost of some additional map distortion. Using a representative sample library of drum sounds, we search for a set of optimal algorithm parameters according to objective measures of map quality and produce a map for the chosen sound corpus. We then conduct a series of qualitative interviews with audio professionals to gain some understanding of the complex situation that is sample library interaction in a music production environment and to gauge initial reactions to the alternative software we developed. Participants’ responses allow us to identify a prevalent method of working with sample libraries, which we codify into a generalized model of the established workflow. The results confirm the need for and interest in alternate interfaces. Although the organization of sounds in the map interface is seen as not easily comprehensible, interview responses confirm the need for and interest in our software.

This work thus presents a functioning proof of principal for the use of SOMs for sound corpus organization. It demonstrates that there is a high interest in such methods. Despite interview participants’ criticism of details, overall feedback is positive. Therefore, further development of the presented work in close exchange with users appears to be very sensible.

## Zusammenfassung

Umfangreiche Sammlungen von Audiodateien — *Klangkorpora* — sind so leicht zugänglich wie nie zuvor. Online verfügbare Sample Libraries und kostengünstige Datenträger haben Speicherplatz als Einschränkung für moderne Musikproduzenten effektiv beseitigt. Zugleich hat jedoch das verfügbare Instrumentarium zur Navigation, Durchsuchung und Organisation dieser zunehmend unüberschaubaren Datensammlungen nicht Schritt gehalten. Das derzeit wohl gängigste Werkzeug, mit dem Produzenten ihre Sample Libraries durchsuchen, sind Dateimanager, welche alphabetisch sortierte Listen von Dateinamen anzeigen.

Die vorliegende Masterarbeit geht dieses Problem praktisch an. Wir implementieren die Self-Organizing Map (SOM), einen bewährten Algorithmus des maschinellen Lernens zur Dimensionsreduktion und Datenvisualisierung, und verwenden diese zur Organisation von Klangkorpora. Wir präsentieren *SOM Browser*, ein Programm mit einer schnellen, visuellen Bedienoberfläche zur Erkundung von Sample Libraries. Es bietet eine Alternative zum etablierten Musikproduktionsprozess, da es den SOM Algorithmus integriert, welcher nach unserem Wissen in keiner kommerziellen Audiosoftware zur Verfügung steht. *SOM Browser* ist eine Standalone-Anwendung, welche einen Klangkorpus komplett überwacht organisiert und eine zweidimensionale Karte der Klänge in einer interaktiven, rasterbasierten Oberfläche präsentiert. Damit können Klangdateien in schneller Abfolge angespielt werden, was es ermöglicht, rasch einen Überblick der analysierten Dateien zu bekommen. Um den von der Karte eingenommenen Platz zu optimieren, erweitern wir den SOM Algorithmus mit einer neuen Methode, welche wir Forced Node Population (FNP) nennen. FNP reduziert unbesiedelte (‘‘leere’’) Bereiche der Karte auf Kosten zusätzlicher Verzerrung. Unter Verwendung einer repräsentativen Sample Library von Schlagzeug- und Perkussionsklängen suchen wir nach einer Reihe von optimalen Parametern des Algorithmus gemäß objektiver Maße für die Qualität der SOM und erzeugen eine Karte für den gewählten Klangkorpus. Danach führen wir qualitative Interviews mit fünf professionellen Produzenten und Künstlern, um ein besseres Verständnis der Komplexität der Interaktion mit Sample Libraries zu gewinnen und um erste Reaktionen auf die von uns entwickelte Softwarealternative zu erfassen. Basierend auf Antworten der Befragten identifizieren wir einen typischen Umgang mit Sample Libraries, kodifizieren diesen und erstellen ein generalisiertes Modell der etablierten Arbeitsmethode. Obwohl die Organisation der Klänge in unserer Bedienoberfläche als nicht leicht nachvollziehbar eingeschätzt wird, bestätigen die erfassten Antworten Bedarf und Interesse an unserer Software.

Diese Arbeit legt somit einen funktionierenden *Proof of Principal* für die Verwendung von SOMs zur Organisation von Klangkorpora vor. Sie zeigt gleichzeitig, dass ein hohes Interesse an solchen Methoden besteht. Trotz Detailkritik der Befragten ist das Feedback von Nutzerseite positiv, sodass die Weiterentwicklung der präsentierten Arbeit in engem Austausch mit Nutzern sehr sinnvoll erscheint.

## Acknowledgements

I would like to express my gratitude to the advisors of this thesis, Prof. Dr. Stefan Weinzierl and Dr. Diemo Schwarz for supporting my work. I am particularly grateful to Diemo Schwarz for allowing me to work with him at IRCAM, where the idea for all of this was born.

I would also like to extend my thanks to Dr. Jochen Steffens for his help in looking over my interview design and to Henrik von Coler for many helpful conversations and tips on how to navigate the process of writing this thesis.

Additionally, I wish to thank Jelle Akkerman for letting me bug him about software architecture and sharing his coding expertise.

None of this would have been possible without my parents and their continuous support, even though finishing this degree has taken a bit of time. I am deeply grateful to you.

Thank you also to my sister Lena, for your empathy as you share my scholastic woes.

Finally, Audri, thank you for always being there. Here's to whatever comes next.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Motivation and Problem Description . . . . .	1
1.2	Previous Work . . . . .	2
1.3	Aims and Objectives . . . . .	3
<b>2</b>	<b>Background</b>	<b>4</b>
2.1	Audio Feature Extraction . . . . .	4
2.1.1	Audio Pre-Processing . . . . .	4
2.1.2	Time Domain Features . . . . .	5
2.1.3	Frequency Domain Features . . . . .	5
2.1.4	Perceptual Features . . . . .	7
2.2	Self-Organizing Map . . . . .	7
2.2.1	Algorithm Definition . . . . .	8
2.2.2	Node Initialization . . . . .	8
2.2.3	Input Data Scaling . . . . .	9
2.2.4	Alternative Learning Rate Factors . . . . .	9
<b>3</b>	<b>Implementation</b>	<b>10</b>
3.1	Groundwork: CataRT Extension . . . . .	10
3.1.1	Functionality . . . . .	12
3.1.2	Code Overview . . . . .	12
3.2	SOM Browser . . . . .	15
3.2.1	Functionality . . . . .	15
3.2.2	Libraries and Frameworks Used . . . . .	17
3.2.3	Application Structure . . . . .	18
3.2.4	Background Processing . . . . .	19
3.2.5	User Interface Components . . . . .	20
3.2.6	Algorithm Extension: Forced Node Population . . . . .	22
<b>4</b>	<b>Evaluation</b>	<b>24</b>
4.1	Sound Corpus Selection . . . . .	24
4.2	Metrics for SOM Analysis . . . . .	25
4.2.1	SOM-Induced Quantization . . . . .	26
4.2.2	Vector-Node Count . . . . .	26
4.2.3	Map Emptiness . . . . .	26
4.2.4	Influence of Forced Node Population . . . . .	27
4.3	Semi-structured User Interviews . . . . .	27
4.3.1	Motivation to Conduct Interviews . . . . .	27
4.3.2	Interview Subject Selection . . . . .	28
4.3.3	Informed Consent Form . . . . .	28
4.3.4	Test Subject Code Design . . . . .	28
4.3.5	Interview Structure . . . . .	28
4.3.6	Question Design . . . . .	29
4.3.7	Selection of Ratings Scales . . . . .	29
4.3.8	Questions Used . . . . .	29

<b>5</b>	<b>Results</b>	<b>31</b>
5.1	SOM Metrics . . . . .	31
5.1.1	Effects of Training Duration . . . . .	31
5.1.2	Effects of $\alpha_{initial}$ . . . . .	31
5.1.3	Effects of Radius Size . . . . .	33
5.1.4	Learning Rate Type Comparison . . . . .	33
5.1.5	Final Map and Influence of Forced Node Population (FNP) . . . . .	35
5.2	Interview Results . . . . .	37
5.2.1	Established Workflow Responses . . . . .	38
5.2.2	SOM Browser Responses . . . . .	39
5.2.3	Workflow Comparison: SOM Browser vs. Established Workflow . . .	42
5.2.4	SOM Browser: Feature Requests . . . . .	42
<b>6</b>	<b>Discussion</b>	<b>44</b>
6.1	Interpreting the Results . . . . .	44
6.2	Strengths, Weaknesses and Limitations . . . . .	46
6.3	Outlook . . . . .	47
6.4	Conclusion . . . . .	47
<b>7</b>	<b>References</b>	<b>49</b>
	<b>Appendices</b>	<b>I</b>
<b>A</b>	<b>Source Code</b>	<b>I</b>
<b>B</b>	<b>Sample Library</b>	<b>I</b>
<b>C</b>	<b>Evaluation Map Data</b>	<b>I</b>
<b>D</b>	<b>Interview Data</b>	<b>I</b>
<b>E</b>	<b>MATLAB Figures</b>	<b>I</b>
	<b>Acronyms</b>	<b>II</b>
	<b>List of Figures</b>	<b>III</b>
	<b>List of Listings</b>	<b>IV</b>
	<b>List of Tables</b>	<b>V</b>



# 1 Introduction

## 1.1 Motivation and Problem Description

Sample libraries are ubiquitous in modern audio production. Affordable storage media and fast computers enable artists, producers and sound designers to work with large collections of audio files - a one terabyte hard drive costs around 50 Euros and can store up to two months of uncompressed audio information in CD quality. At the same time, the Internet is full of large sound corpora for anyone to download. This is especially true for collections of drum sounds, which are offered for sale by many companies or traded (sometimes without considerations for any applicable copyright) on various web forums. Consequently, many users of digital audio software amass vast collections of these “sample packs”, often without ever listening to or being familiar with all of their contents, as they encounter the *paradox of choice* (Schwartz, 2004). The technological development enabling this abundance of sound samples also presents new and serious challenges to the efficient use of them: how can such large numbers of files be searched, organized, compared, and presented to the user?

The present thesis explores a practical approach to the problem of sound corpus organization, while focussing primarily on drum samples as used by electronic music producers. *Sound corpus* in this work denotes a collection of sound files — other commonly encountered terms for such collections are sample library, pack, kit or bundle. Currently, arguably the most common way in which producers search through their sample libraries is by using some kind of file browser (built into their computer’s Operating System (OS) or their Digital Audio Workstation (DAW) of choice). This browser presents a list view of all audio files in the current folder and is typically sorted alphabetically, chronologically or by some other criterion that is almost definitely not directly related to the sonic content of the file.

At the same time, advances in the field of Music Information Retrieval (MIR) have enabled researchers and developers to extract descriptive information about the contents of a digital audio file for decades now, making it possible to present much more relevant data to software users. Commercial software tools incorporating Audio Content Analysis (ACA) are slowly becoming available (see Section 1.2 below). Still, the problem of data visualization persists - how can all these additional dimensions of information be displayed in a way that improves users’ workflows?

In order to find an alternative organizational method to the name-based, categorical file browser interface described above, we turn to Machine Learning (ML), a field that deals with pattern recognition and classification tasks. The *Self-Organizing Map (SOM)* algorithm, first introduced by Teuvo Kohonen (see Section 2.2, Kohonen (1990)), is a machine learning algorithm that performs dimensionality reduction on a set of higher-dimensional input data and can at the same time be used for data visualization, as its output is often two-dimensional and can be shown as a regular grid structure. Besides being an established algorithm that has been extensively evaluated and used in various applications (see for example Kohonen (1990, p.1476) for an overview), it offers a number of further advantages that informed our decision to employ it in this work. It can visualize high-dimensional data sets, while training itself completely unsupervised. It can organize unnamed audio files, meaning it can offer up some proposal for a structure without any metadata. Lastly, the SOM is based on a grid layout, which is an influential, common

structure in electronic music hardware and music technology in general - grids are ever-present in electronic music studios (Adeney and Brown, 2009).

## 1.2 Previous Work

The topic of using audio descriptors (see Section 2.1 for further details) for the organization of sounds in two-dimensional interfaces, often displayed as scatter plots, has been explored in a variety of previous research. Particularly relevant for the present work is the software *CataRT* by Schwarz et al. (2006). It allows realtime corpus-based concatenative synthesis and offers the user an interface to explore the loaded sound corpus interactively. Two selectable audio descriptors make up the axes along which sounds are plotted as circles, while two more descriptors can be mapped to circle color and size. At the same time, more descriptors are already calculated, but cannot be displayed simultaneously, which informs our decision to extend *CataRT* with the option of using a two-dimensional SOM as the interface.

Coleman (2007) also uses feature extraction to create a scatter plot interface of a *personal sample library*, which directly relates to our use case as described in Section 1.1 above. The author also implements a way to filter samples by feature values using *dynamic queries* through realtime interaction with user interface elements such as sliders and check boxes.

An early example of using SOMs for audio data is Cosi et al. (1994), who use the algorithm to classify classical musical instruments. They are able to distinguish between twelve instrumental timbres and also show that new timbres not used during training can be classified by the map, which will place the unknown timbre close to the most similar training timbre.

Dealing explicitly with drum samples, Pampalk et al. (2004) use a combination of one- and two-dimensional SOMs. The one-dimensional variant of the algorithm is used to create a hierarchical sample structure, while a two-dimensional SOM is used for visualization of the samples.

Heise et al. (2008) combine a SOM-based interface with surround sound reproduction for an interesting and novel approach to sound corpus exploration with their *SoundTorch* software. The authors let the user carry the eponymous *SoundTorch*, which is used to illuminate parts of a map interface onto which sounds are placed. Illuminated sounds are played back on a multichannel loudspeaker system, with their position in the interface determining their spatial playback position.

Fried et al. (2014) are employing a different approach to visualization and dimensionality reduction of high-dimensional descriptor spaces with their *AudioQuilt* software, which relies on user input for its similarity measure. Their approach achieves a 1-to-1 mapping between sound samples and grid locations at the cost of not functioning completely unsupervised.

Finally, we can also give a couple of examples of relevant software that exist outside of academia. The *Infinite Drum Machine* by McDonald and Tan (2019) uses the t-SNE algorithm (Maaten and Hinton, 2008) to create a drum machine with a two-dimensional sound selection interface that resembles a point cloud. A commercial software tool for organizing and maintaining sound corpora that has been around for several years now is *Audio Finder* (Iced Audio, 2019), which does not employ machine learning at all, instead

offering a large variety of metadata and cataloging tools, as well as some spectral analysis features. More recently, commercially available audio software that incorporates machine learning is starting to appear. Two examples of this are *Sononym* (Nielsen, 2019) and *Atlas* (Algonaut, 2019). Similarly to *Audio Finder*, *Sononym* deals with sample library organization and offers a ML-based “similarity search”, while *Atlas* is a drum sampler plug-in that also features a sound map.

### 1.3 Aims and Objectives

The aim of this thesis is to implement the *SOM* algorithm from scratch and use it to build a software interface for sound corpus organization and exploration, as there is currently no readily available audio software with this functionality. To this end, we incorporate the algorithm into *CataRT*, an existing software for corpus-based concatenative synthesis (CBCS) (Schwarz et al., 2006). We also develop a larger standalone application called *SOM Browser* that uses the SOM as an alternative interface for the exploration of a folder of samples, as this is something that no commercially available software offers. Subsequently, an overview of the effects of the different controllable parameters of the algorithm on the produced SOM is given and an ideal map for a representative sound corpus of drum samples is shown. Finally, we present user feedback to *SOM Browser* gathered from interviews with five audio professionals. The main question we aim to explore during these interviews is whether our software can offer a viable alternative to established workflows or if our proposed alternative interface is perhaps too abstract to be considered useful.

## 2 Background

The following chapter intends to provide a methodological background for two key concepts underlying the work presented in this thesis, namely Audio Feature Extraction and the Self-Organizing Map.

### 2.1 Audio Feature Extraction

Audio Feature Extraction is the process of deriving *features* from a digital audio signal. A feature represents some sort of descriptive information about the audio data. According to Lerch (2012), this extraction process serves a dual purpose; that of dimensionality reduction as well as a more meaningful representation. A large variety of features for different purposes have been developed (refer to Peeters (2004) for an extensive list, as well as Lerch (2012) for an in-depth look at the topic). The following subsections introduce the features used in this work, starting with the pre-processing required to prepare an audio signal for feature extraction and then moving into individual definitions for each feature. The equations presented here are based on the formal definitions given by Lerch (2012) along with the computational implementations of the features in the *Meyda* library for feature extraction in JavaScript (Rawlinson et al., 2015), which in turn adapted the *yaafe* library for Python (Mathieu et al., 2010).

#### 2.1.1 Audio Pre-Processing

Consider a digital audio signal of the form  $x[n]$ , where  $n$  denotes the sample index and  $x[n]$  the value of the individual sample at that index.

**Normalization** In order to have a standardized maximum amplitude of 1 across all audio signals, they are normalized such that

$$x_{norm}[n] = \frac{x[n]}{\max x[n]}. \quad (1)$$

**Mono Conversion** Spatial information, as contained in an audio file with more than one channel, is not deemed necessary in the presented work. For this reason, all audio signals are converted to mono by taking the average of all channels.

**Frame-Based Feature Extraction** Rather than performing feature extraction on the entirety of the audio signal, it is common practice to divide the signal into smaller chunks or *frames*, typically consisting of some  $2^n$  samples (512, 1024, 2048 are often found values). The resulting feature values for each frame form a trajectory of the feature's evolution over time, which can either be used as such or can be averaged. In this work, audio signals are divided into frames with a length of 512 samples. In order to avoid computational errors (such as Not a Number (NaN) in JavaScript) during potentially silent portions of the audio signal, frames with  $v_{RMS} < -60$  dBFS (see Root Mean Square (RMS) definition below) are omitted from the feature extraction. The following equations define each feature for a single frame.

### 2.1.2 Time Domain Features

Time domain features are features derived directly from the discrete-time signal  $x[n]$ .

**Duration** The overall duration of the signal  $x[n]$  in seconds:

$$v_{DUR} = \frac{n}{fs}s, \quad (2)$$

where  $n$  is the number of samples and  $fs$  is the sampling rate.

**Root Mean Square (RMS)** measures the power of a signal (Lerch, 2012, p.73f). It describes sound intensity and is sometimes used as a simple measure for loudness (Rawlinson et al., 2019a) that does not take the nonlinearity of human hearing into account (Fletcher and Munson, 1933). It is calculated for an audio frame  $x[n]$  consisting of  $n$  samples such that

$$v_{RMS} = \sqrt{\frac{\sum_{i=1}^n x(i)^2}{n}}. \quad (3)$$

**Zero-Crossing Rate (ZCR)** represents the rate of the number of sign changes in a signal. It can be used as a measure of the tonalness of a sound (Lykartsis, 2014) and as a simple pitch detection method for monophonic signals (de la Cuadra, 2019). It is defined as

$$v_{ZCR} = \frac{1}{2 \cdot n} \sum_{i=1}^n |sgn[x(i)] - sgn[x(i-1)]|. \quad (4)$$

### 2.1.3 Frequency Domain Features

Frequency domain features or *spectral* features are derived from the discrete complex spectrum  $X(k)$ , where  $k$  refers to the frequency bin number.  $X(k)$  is calculated from  $x[n]$  by performing a Fast Fourier Transform (FFT) (for information on the Fourier transform, refer to any signal processing textbook, such as Oppenheim and Schaffer (2014)).

**Spectral Centroid** is a measure of the center of gravity of a spectrum. A higher value indicates a brighter, sharper sound (Lerch, 2012). The spectral centroid is defined as

$$v_{SC} = \frac{\sum_{k=0}^{N_{FFT}/2-1} k \cdot |X(k)|^2}{\sum_{k=0}^{N_{FFT}/2-1} |X(k)|^2}. \quad (5)$$

**Spectral Flatness** is a measure for the tonality or noisiness of a signal, defined as the ratio of the geometric and arithmetic means of its magnitude spectrum. Higher values

indicate a flatter (and therefore noisier) spectrum, whereas lower values point towards more tonal spectral content. It is defined as

$$v_{SFL} = \frac{N_{FFT}^{1/2} \sqrt{\prod_{k=0}^{N_{FFT}/2-1} |X(k)|}}{(2/N_{FFT}) \cdot \sum_{k=0}^{N_{FFT}/2-1} |X(k)|}. \quad (6)$$

**Spectral Kurtosis** indicates whether a given magnitude spectrum's distribution is similar to a Gaussian distribution. Negative values result from a flatter distribution, whereas positive values indicate a peakier distribution. A Gaussian distribution would result in a value of 0. Spectral Kurtosis is defined as

$$v_{SKU} = \frac{2 \sum_{k=0}^{N_{FFT}/2-1} (|X(k)| - \mu_{|X|})^4}{N_{FFT} \cdot \sigma_{|X|}^4} - 3, \quad (7)$$

where  $\mu_{|X|}$  represents the mean and  $\sigma_{|X|}$  the standard deviation of the magnitude spectrum  $|X|$ .

**Spectral Skewness** assesses the symmetry of a magnitude spectrum distribution. It is defined as

$$v_{SSK} = \frac{2 \sum_{k=0}^{N_{FFT}/2-1} (|X(k)| - \mu_{|X|})^3}{N_{FFT} \cdot \sigma_{|X|}^3}. \quad (8)$$

**Spectral Slope** represents a measure of how sloped or inclined a given spectral distribution is. The spectral slope is calculated using a linear regression of the magnitude spectrum such that

$$v_{SSL} = \frac{\sum_{k=0}^{N_{FFT}/2-1} (k - \mu_k)(|X(k)| - \mu_{|X|})}{\sum_{k=0}^{N_{FFT}/2-1} (k - \mu_k)^2}. \quad (9)$$

**Spectral Spread** is a descriptor of the concentration of a magnitude spectrum around the Spectral Centroid and assesses the corresponding signal's bandwidth. It is defined as

$$v_{SSP} = \frac{\sum_{k=0}^{N_{FFT}/2-1} (k - v_{SC})^2 \cdot |X(k)|^2}{\sum_{k=0}^{N_{FFT}/2-1} |X(k)|^2}. \quad (10)$$

**Spectral Rolloff** measures the bandwidth of a given signal by calculating that frequency bin below which lie  $\kappa$  percent of the sum of magnitudes of  $X(k)$ . Common values for  $\kappa$  are 0.85, 0.95 (Lerch, 2012) or 0.99 (Rawlinson et al., 2019a). It is defined as

$$v_{SR} = i \left| \sum_{k=0}^i |X(k)| = \kappa \cdot \sum_{k=0}^{N_{FFT}/2-1} |X(k)| \right|. \quad (11)$$

#### 2.1.4 Perceptual Features

Both the time and frequency domain features introduced above are derived from raw audio samples without taking into account any concept of human sound perception. Perceptual features incorporate some sort of model that approximates this perception. While only a single perceptual feature is used in this work, more do exist (see Peeters (2004) for a list of some of them).

**Total Loudness** represents an algorithmic approximation of the human perception of a signal’s loudness based on Moore et al. (1997), which uses the Bark scale as introduced by Zwicker (1961). The Total Loudness is the sum of all 24 bands’ specific loudness coefficients, defined by Peeters (2004) as

$$v_{TL} = \sum_{i=1}^{24} v_{SL}(i), \quad (12)$$

where

$$v_{SL}(i) = E(i)^{0.23} \quad (13)$$

is the specific loudness of each Bark band (see Moore et al. (1997) for further details).

## 2.2 Self-Organizing Map

The *self-organizing map* (SOM) is a machine learning algorithm for dimensionality reduction, visualization and analysis of higher-dimensional data. Sometimes also referred to as *Kohonen map* or *network*, it was introduced in 1981 by Teuvo Kohonen (Kohonen, 1990). The SOM is a variant of an *artificial neural network* that uses an unsupervised, competitive learning process to map a set of higher-dimensional observations (the *input vectors*) onto a regular, often two-dimensional grid or *map* of *neurons* or *nodes* that is easy to visualize. The SOM can be regarded as a nonlinear generalization of a principal component analysis (PCA) (Yin, 2007) or as a quantization of the input data, with the nodes along the map functioning as pointers into that higher-dimensional space. Each node has a position on the lower-dimensional grid as well as an associated position in the input space, which takes the form of a  $n$ -dimensional weight vector  $m = [m_1, \dots, m_n]$ , where  $n$  is the number of dimensions of the input vectors. Nodes that are in close proximity to each other on the SOM will also have similar weight vectors (Vesanto et al., 2000), although the inverse (neighboring positions in the input space also mapping to neighboring nodes) is not necessarily true (Bauer et al., 1996).

For an in-depth look at the algorithm, its variants and applications, as well as an extensive survey of research on SOMs, the avid reader is referred to Kohonen (2001).

### 2.2.1 Algorithm Definition

The following definition is based on Kohonen (1990), Kohonen (2005), Kohonen and Honkela (2007) and Bauer et al. (1996).

Consider a space of input data in the form of  $n$ -dimensional vectors  $x \in \mathbb{R}^n$  and an ordered set of nodes or model vectors  $m_i \in \mathbb{R}^n$ . A vector  $x(t)$  is mapped to that node  $m_c$  with the shortest Euclidean distance from it:

$$\|x(t) - m_c\| \leq \|x(t) - m_i\| \forall i. \quad (14)$$

This “winning” node  $m_c$  is referred to as the Best Matching Unit (BMU) for  $x(t)$ .

During the learning or adaptation phase of the algorithm, all nodes  $m_i$  are adjusted by a recursive regression process

$$m_i(t+1) = m_i(t) + h_{c(x),i}(x(t) - m_i(t)), \quad (15)$$

where  $t$  is the index of the current regression step,  $x(t)$  is an input vector chosen randomly from the input data at this step,  $c$  is the index of the BMU for the current input vector  $x(t)$  according to Equation 14 and  $h_{c(x),i}$  represents a so-called *neighborhood function*. The name-giving *neighborhood* is a subset  $N_c$  of nodes centered on  $m_c$ . At each learning step  $t$ , those nodes that are within  $N_c$  will be adjusted, whereas those outside of it will not. The reason for employing such a neighborhood function is so that the nodes “doing the learning are not affected independently of each other” (Kohonen, 1990, p.1467) and “the topography of the map is ensured” (Bauer et al., 1996, p.5). At its most basic, the neighborhood function is a decreasing distance function between neurons  $m_i$  and  $m_c$ . Its most common form, which is also employed in this work, is that of a Gaussian function with its peak at  $m_c$  such that

$$h_{c(x),i} = \alpha(t) \exp\left(-\frac{\|r_i - r_c\|^2}{2\sigma^2(t)}\right). \quad (16)$$

Here,  $\alpha$  denotes a learning rate factor or adaptation “gain control”  $0 < \alpha(t) < 1$ , which decreases over the course of the regression,  $r_i \in \mathbb{R}^2$  and  $r_c \in \mathbb{R}^2$  are the locations of  $m_i$  and  $m_c$  on the SOM grid (the lower-dimensional output map, not the input space!), and  $\sigma(t)$  is the width of the neighborhood function, which again decreases as the regression step index increases.

### 2.2.2 Node Initialization

Because of the iterative nature of the SOM algorithm, its outcome depends on the initial positions chosen for the nodes. The method implemented in this work uses random initialization, meaning the starting positions of the nodes are chosen randomly from within the bounds of the input space. An often employed alternative approach is to first perform a Principal Component Analysis (PCA) on the input data, select the largest  $d$  components, where  $d$  is the number of desired output dimensions for the SOM, and then distribute the nodes at equidistant intervals along those component vectors.



### 2.2.3 Input Data Scaling

Some consideration should be given to the dynamic range of the input data across its different dimensions. Are the dimensional ranges comparable in their limits? What about their variance? There does not appear to exist a clear consensus across the literature on whether or not normalization of input data is strictly necessary (Vesanto et al. (2000, p.34), Kohonen (1990, p.1470), Kohonen and Honkela (2007)).

Because the range of the data derived from the audio feature analysis used in this work varies considerably between features, the data for feature  $n$  is rescaled to have unit variance by dividing by the feature's standard deviation  $\sigma_n$ :

$$x_n = \frac{x_n}{\sigma_n}. \quad (17)$$

### 2.2.4 Alternative Learning Rate Factors

**Linear** The traditional SOM algorithm uses a learning rate factor  $\alpha$  that decreases linearly as a function of the regression step  $t$ :

$$\alpha(t) = \alpha_0 \left( \frac{1-t}{T} \right), \quad (18)$$

where  $\alpha_0$  is the initially chosen learning rate and  $T$  is the total training length or number of regression steps.

Two other approaches to the decreasing learning rate factor were implemented in this work:

**Inverse** The first is a reciprocally decreasing function where

$$\alpha(t) = \frac{\alpha_0}{\left( \frac{1+100t}{T} \right)}. \quad (19)$$

**Bauer-Der-Herrmann Algorithm (BDH)** The second alternative approach is that of an adaptive local learning rate as developed by Bauer et al. (1996) (referred to hereafter as BDH, also see Merenyi et al. (2007)):

$$\alpha(t) = \alpha_0 \left( \frac{1}{\Delta t_c} \left( \frac{1}{|x(t) - m_c|^n} \right) \right)^m, \quad (20)$$

where  $\Delta t_c$  represents the time since the current BMU for the current input vector was last selected as a BMU for any vector and  $m$  is a newly introduced, free control parameter. For a more complete review of the uses of this algorithm, the reader is referred to the original paper (Bauer et al., 1996) as well as Merenyi et al. (2007).

### 3 Implementation

After some methodological background information was given in the previous chapter, the following sections aim to explain how the SOM algorithm was implemented in JavaScript. First, a smaller program was built to extend the existing software *CataRT*. Then a second, more fully fledged application called *SOM Browser* was developed. For both of these programs, we take a look at their functionality and features, give an overview of the code and program structure, and explain some concepts and considerations that were important for the development process.

#### 3.1 Groundwork: CataRT Extension

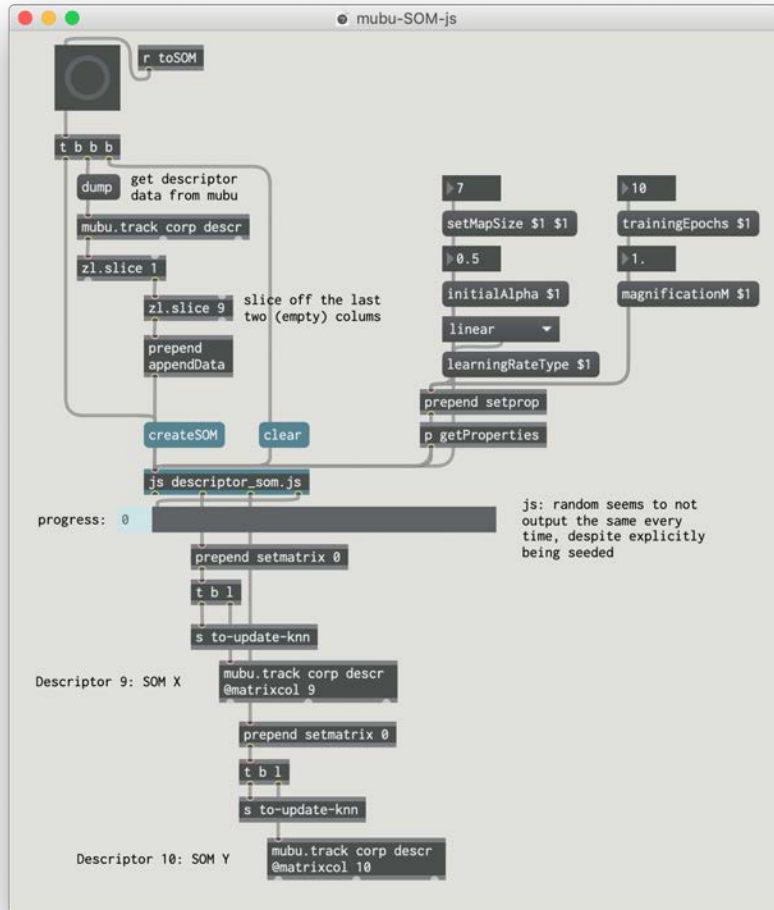


Figure 1: mubu-SOM-js

For the purpose of laying the groundwork for a bigger standalone application (see section 3.2), a proof-of-concept implementation of the core SOM algorithm was written in JavaScript to serve as an extension to the *MuBu For Max* software package (Schnell

et al., 2019a,b) for the visual programming language Max (Cycling '74, 2019). *MuBu For Max* was developed by the Sound, Music, Movement, Interaction Team (ISMM) at Institut de recherche et coordination acoustique/musique (IRCAM) (Schnell et al., 2009). It contains the *catart-by-mubu* patch for realtime interactive corpus-based concatenative synthesis based on the original *CataRT* software (Schwarz et al., 2006). The developed extension is a Max patch called *mubu-SOM-js* (see Figure 1) and can be found as part of the digital submission of this thesis in the directory `dev/mubu-som-js`.

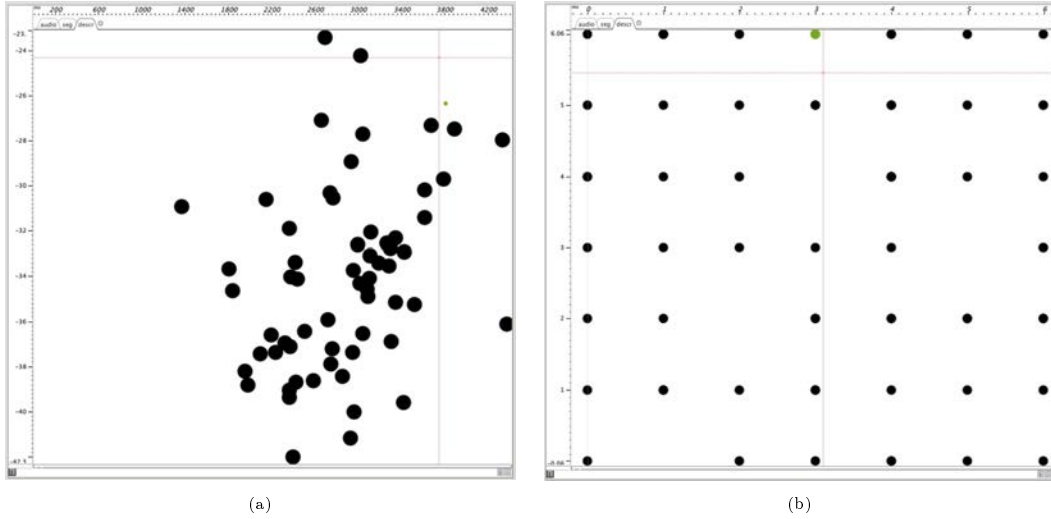


Figure 2: *CataRT* display of a corpus without SOM (2a, X axis shows spectral centroid, Y axis shows loudness) and with SOM extension (2b). Each circle represents a sample.

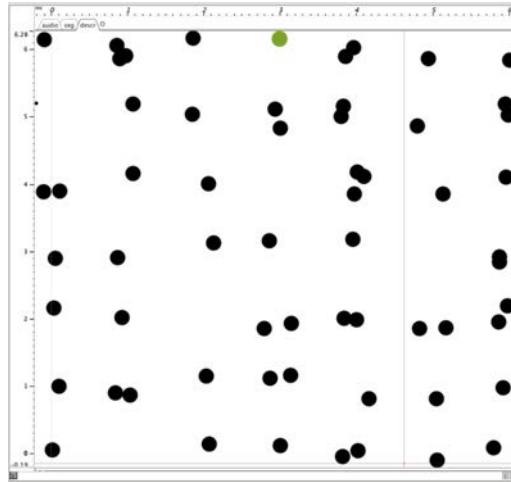


Figure 3: *CataRT* display of a corpus with SOM extension and added noise to spatially differentiate samples assigned to the same node. Each circle represents a sample.

*Catart-by-mubu* uses a two-dimensional scatter plot interface in which the user can select samples or grains from the loaded audio corpus (see Figure 2). The spatial position of these sounds in the interface is determined by two audio features, representing the horizontal and vertical axes, that can be selected by the user. The implemented SOM extension

gives users the option to choose a two-dimensional SOM for the spatial organization of the corpus. This augments the interface in three ways: all analyzed audio features can be taken into account for the spatial positioning (as opposed to just two at a time), more of the available interface space is used and additionally the sounds are spaced in a more even fashion (see Figure 2).

### 3.1.1 Functionality

*Mubu-SOM-js* offers the user simple controls to influence the produced SOM. These can be set by sending the messages outlined in Table 1 to the `[js_descriptor_som.js]` object.

Message	Type	Description	Example
<code>createSOM</code>	n/a	Initiates SOM calculation.	<code>createSOM</code>
<code>setMapSize \$1 \$1</code>	Float	Sets size of map.	<code>setMapSize 7 7</code>
<code>trainingEpochs \$1</code>	Int	Defines the length of the training in epochs. One epoch corresponds to $n$ iterations of the training algorithm (see section 2.2.1), where $n$ is the number of samples in the corpus.	<code>trainingEpochs 30</code>
<code>initialAlpha \$1</code>	Float	Sets the starting value for the learning rate factor $\alpha$ .	<code>initialAlpha 0.5</code>
<code>learningRateType \$1</code>	String	Sets the learning rate type (see section 2.2.4). It expects a string that is either <code>'linear'</code> , <code>'inverse'</code> or <code>'BDH'</code> .	<code>learningRateType 'linear'</code>
<code>magnificationM \$1</code>	Float	Sets the magnification control factor $m$ (see section 2.2.4). Only applies when <code>learningRateType === 'BDH'</code> .	<code>magnificationM 0.02</code>

Table 1: mubu-SOM-js: Messages for algorithm control

### 3.1.2 Code Overview

The core of the *mubu-SOM-js* Max patch is a JavaScript program (see the file `dev/mubu-som-js/descriptor_som.js`). The choice of programming language was determined by the fact that *CataRT* is a Max patch and JavaScript (via the built-in `[js]` object) can be used to script most aspects of the Max environment. This JavaScript version of the SOM is in some ways a port from a first MATLAB implementation of the algorithm that was developed by the author during an internship at IRCAM in the fall of 2017. Some aspects of the structure of the presented program are based on the SOM Toolbox that was developed at Helsinki University of Technology by Vesanto et al. (2000).

The flow of the script is encapsulated in `createSOM()`. This function calls all other important functions that make up the program, as can be seen in Listing 1.

```

34  function createSOM()
35  {
36      normalizeData();
37      initializeMap();
38      trainMap();
39  }

```

Listing 1: dev/mubu-som-js/descriptor\_som.js: createSOM()

After data normalization and map initialization, `trainMap()` is called, which executes the training procedure by repeatedly calling the function `training()` in an asynchronous background process (see Listing 2). For each step of the training phase, all calculations happen inside `trainingStep()`. The most important part, the updating of node positions on each iteration, is shown in Listing 3.

```

203  function training()
204  {
205      if (t < trainingLength)
206      {
207          trainingStep(t, trainingLength, rStep, alpha, winTimeStamp);
208          // Progress percentage on outlet 4
209          outlet(3, math.ceil(100 * (t / trainingLength)));
210          t++;
211      }
212      else
213      {
214          post('Training done.\n');
215          findBestMatches();
216          outputDataCoordinatesOnMap();
217          arguments.callee.task.cancel();
218      }
219  }

```

Listing 2: dev/mubu-som-js/descriptor\_som.js: training()

```

307      // For each neuron, get neighborhood function and update its position.
308      neurons = neurons.map(function (neuron, index) {
309          // Gaussian neighborhood function
310          var h = alpha * math.exp(-(math.square(distances[index][bmu])
311                                     / (2 * math.square(r))));
312          return math.subtract(neuron, math.multiply(h, differences[index]));
313      });

```

Listing 3: dev/mubu-som-js/descriptor\_som.js: Neuron positions are updated in `trainingStep()`. This Listing implements equation 15 from Section 2.2.1.

After the training phase is finished, the final map is populated by iterating over all vectors and finding their corresponding best matching units (meaning that node which is

closest), as can be seen in Listing 4. A problem arises when multiple vectors share the same node as their best matching unit; these vectors will have the same exact position in the interface with no option to differentiate between them and no indication of how many vectors reside in that location. To circumvent this issue, a small amount of random noise is added to each vector's position. This creates clusters around the exact node position and allows for the individual circles to be selected. An example of a map with added noise can be found in Figure 3.

```

316 function findBestMatches()
317 {
318   bestMatches = normalizedData.map(function (vector) {
319     var differences = [];
320     var distancesFromVector = [];
321
322     // Subtract chosen vector from each neuron / map unit, then calculate that
323     // difference vector's magnitude.
324     // In other words, calculate the Euclidean distance between each neuron and
325     // the chosen vector.
326     for (var n = 0; n < neuronCount; n++)
327     {
328       distancesFromVector.push(math.norm(math.subtract(neurons[n], vector)));
329     }
330     // Find best matching unit's distance and index:
331     var bmuDistance = math.min(distancesFromVector);
332     var bmu = distancesFromVector.indexOf(bmuDistance);
333     return [bmu, bmuDistance];
334   });
335 }

```

Listing 4: `dev/mubu-som-js/descriptor_som.js`: BMUs for each vector are identified in `findBestMatches()`. This Listing implements equation 14 from Section 2.2.1.

## 3.2 SOM Browser

The majority of the work for this thesis consisted of the development of a standalone application for sample library exploration which we call *SOM Browser*. A screenshot of the program can be seen in Figure 4. All source code for the application is located in the digital submission package of this thesis in the directory `dev/som-browser`.

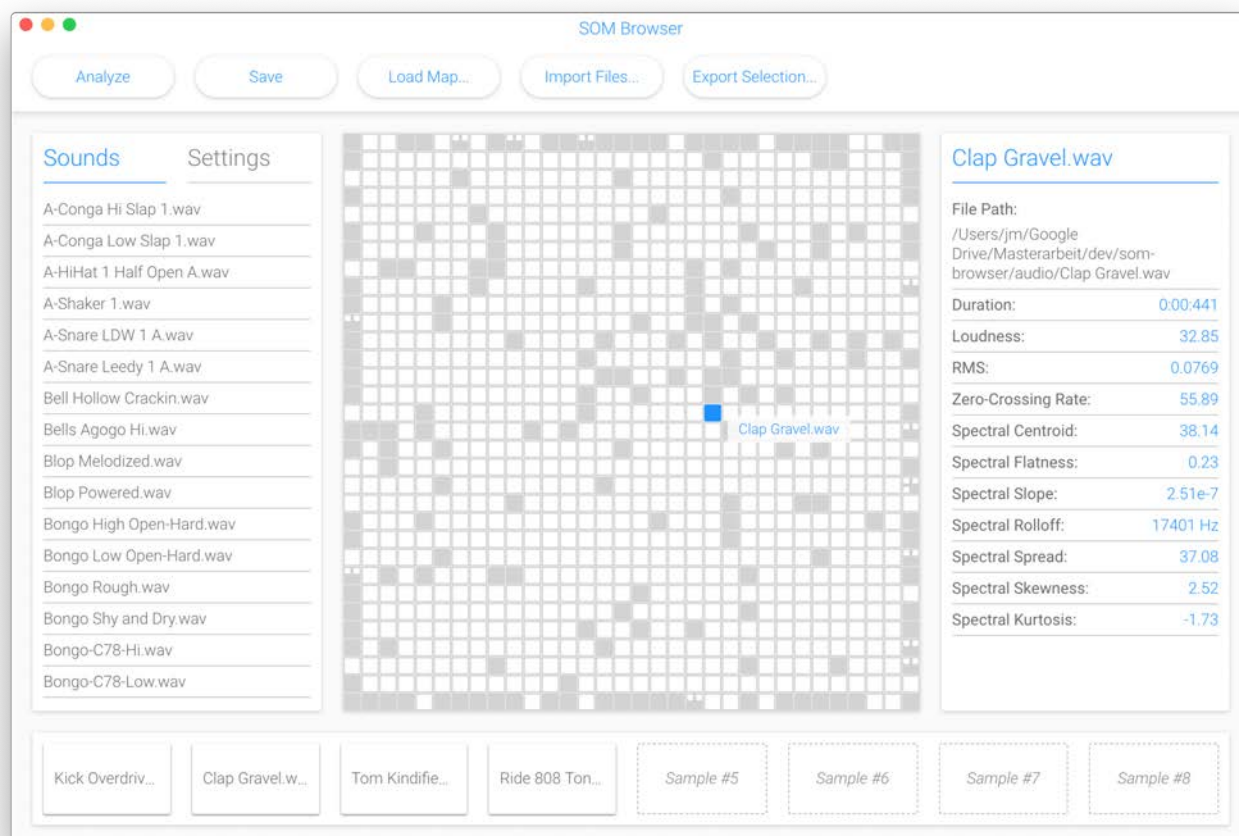
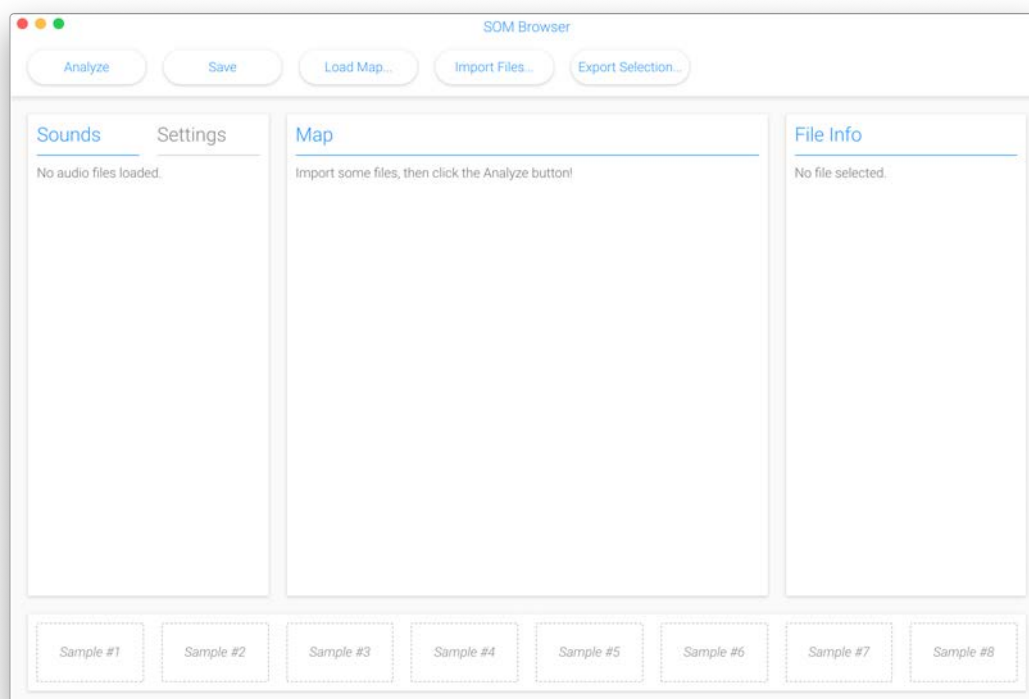


Figure 4: *SOM Browser*

*SOM Browser* offers users an alternative interface for the interaction with a folder of audio samples. Instead of the traditional file browser interface consisting of an alphabetical list of file names, the presented application offers a spatial map layout of the samples, with the aim of allowing users a more direct interaction and giving them a quicker overview of the sounds.

### 3.2.1 Functionality

**Loading Audio Files** When launching *SOM Browser*, the application opens with no sounds or map loaded (see Figure 5). In order to create a map of a collection of sound files, the user can go to the menu bar at the top of the window and click the “*Import Files...*” button to load several audio files.

Figure 5: *SOM Browser* without audio files loaded

**Calculating a Map** Once files are selected, the *Sounds* list on the left side of the application will be populated. Next, by clicking “*Analyze*”, the program will start to analyze the audio files in the background, first extracting audio features (see section 2.1) and then using this information to calculate a SOM using default settings. Alternatively, some SOM parameters can be altered by selecting the *Settings* field next to *Sounds* and adjusting the exposed parameters. Depending on the number of audio files to analyze and the selected training duration, the algorithm will take a while to process. Training progress is indicated as a percentage in the central *Map* panel.

**Map Interaction** Upon completion of the SOM calculation, the *Map* panel will be populated by a grid of white and grey squares. Each white square represents a single sound file. All files are loaded into the computer’s Random-Access Memory (RAM) for quick access. Grey squares are empty nodes, meaning nodes to which no sound files were assigned. Sounds can be played by clicking on the white squares. They can also be played immediately by holding down the Shift key and hovering over them. This allows the user a very fast audition process and makes it possible to play back many files in fast succession, enabling very quick browsing of all loaded audio files. When hovering over a square, the corresponding file name is shown next to the mouse cursor. More detailed information about the file, including its full path, duration and audio feature values can be found in the *FileInfo* panel to the right of the map.



**Selecting and Exporting Favorites** The bottom of the window is taken up by the *Favorites* bar. If a sample is found on the map that the user would like to save for further usage, they can drag the square from the map down into one of the slots labeled “*Sample #1 - #8*”. Samples can also be played from the *Favorites* bar by clicking on them. If the user is satisfied with their selection of samples, they can export the selected *Favorites* (e.g. for further usage in a DAW) by clicking on “*Export Selection*” in the top menu bar. This will open a file dialog window to select a location where the files should be stored.

**Saving and Loading Maps** *SOM Browser* also offers the ability to save entire maps to disk for recall in a later session or import previously stored maps by clicking on the menu bar buttons “*Save*” and “*Load Map*”.

### 3.2.2 Libraries and Frameworks Used

Although a desktop application, *SOM Browser* was built entirely using web technologies, most importantly JavaScript, in order to support multiple OSs with minimal effort. A vast variety of libraries and frameworks are available to use for all aspects of the development process. The following paragraphs outline the tools chosen for this application and their benefits.

**Electron** “is an open source library developed by GitHub for building cross-platform desktop applications with HTML, CSS, and JavaScript. Electron accomplishes this by combining Chromium and Node.js into a single runtime and apps can be packaged for Mac, Windows, and Linux” (GitHub, 2019). It offers a variety of Application Programming Interfaces (APIs) to offer native menus, interact with the file system and more. Its `ipcMain` and `ipcRenderer` APIs are used for asynchronous communication between the Graphical User Interface (GUI) and processes running in the background.

**React** is a JavaScript library for building user interfaces (Facebook, 2019). It breaks the GUI into smaller, self-contained units called *components* that can be independently updated and rendered.

**Web Audio API** enables audio processing and synthesis in (web) applications (World Wide Web Consortium (W3C), 2019). The use of this API makes it possible to write all audio processing code for the presented work in JavaScript. Its core concept is the *audio routing graph*, made up of *audio nodes* (simple building blocks such as an oscillator or a recording). This graph connects sources to other nodes (e.g. effects or filters) and finally to an output destination.

**Meyda** “is a Javascript audio feature extraction library. Meyda supports both offline feature extraction as well as real-time feature extraction using the Web Audio API” (Rawlinson et al., 2019b). Its effectiveness has been validated by researchers at Queen Mary University (“Meyda [...] provide[s] excellent real time feature extraction tools”, Moffat et al. (2015)).

### 3.2.3 Application Structure

*SOM Browser* is a *stateful* application, meaning it is designed to remember user interactions, save its internal data (the *state* of the application) between interaction steps and allow the storing of state data between sessions.

**System States** Before the start of the development process, a set of system states was designed to represent the states through which the application is supposed to progress. These states and their order are shown in Figure 6, giving an abstract overview of the flow of the program. Each panel represents a state and consists of a title (shown in capitalized words at the top, e.g. *Map Created*), a method describing a state transition (underscored and in lower case, e.g. *show map*) and the next state to transition into (bottom right, marked by an arrow, e.g. *→ File Audition*).

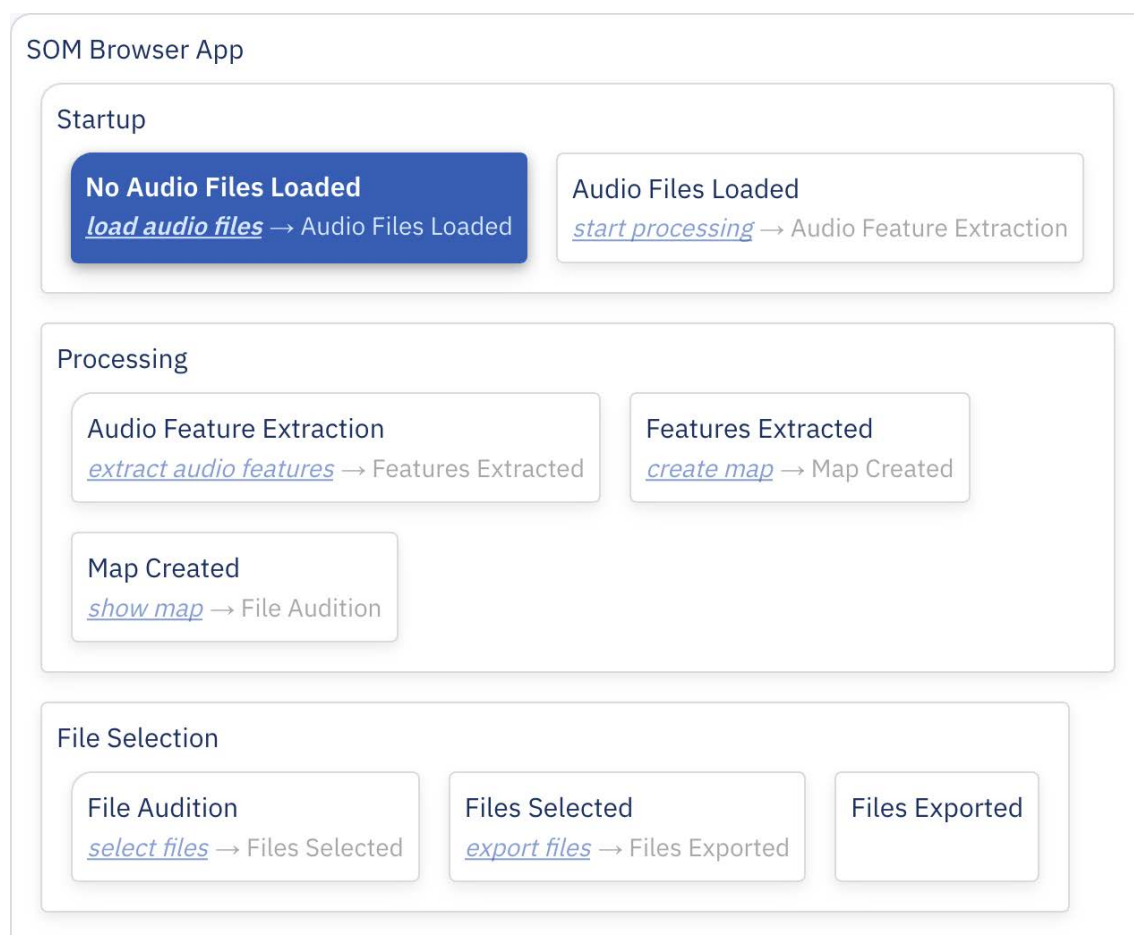


Figure 6: *SOM Browser*: Mock-up outlining system states

**Code Overview** *SOM Browser* was developed using the *git* version control system (Torvalds and Hamano, 2019) in a repository on GitHub <sup>1</sup>. The very basic structure of the application, in particular the way in which the Electron and React frameworks interact, is based on a boilerplate project by Phillip Barbiero (Barbiero, 2017).

The entry point of any Electron application is the `main.js` file, which in the presented work can be found in `dev/som-browser/main.js`. This file creates an instance of the `BrowserWindow` class called `mainWindow` that serves as the single visible application window. `mainWindow` then loads `dev/som-browser/src/index.js`, which imports the React library and uses the React function `render()` to create the `<App />` component, which is defined in `dev/som-browser/src/components/App.js`. It serves as a container for the rest of the application logic and the entire GUI (see Listing 7 and Section 3.2.5 for more details). From here, the structure of the source files branches out into the individual GUI elements in `dev/som-browser/src/components/` and a set of files in `dev/som-browser/src/background/` containing the code for audio feature extraction and SOM calculation.

### 3.2.4 Background Processing

Both audio feature extraction and SOM calculation are processing intensive tasks, therefore it was clear from the beginning of the development stage that these parts of the application must be separated from the GUI that the user interacts with. While it is not possible to build a truly multithreaded application (due to the fact that the fundamental Node.js framework is single threaded), one can create separate processes to run different tasks asynchronously. This is done by creating multiple `BrowserWindow` instances, as each window is running in its own process. These windows can have their `show` flag set to `false` in order to hide them, thereby creating an invisible window for a background process.

*SOM Browser* initiates two consecutive background processes when the user clicks on “Analyze”, one for feature extraction and one that runs the SOM algorithm. This is handled by the function `handleAnalyzeClick()` (found in `dev/som-browser/src/components/App.js`), which passes the necessary data to these background processes by calling `processFiles(files)` and `createSOM(files, settings)` (see Listing 5). Note the chaining of commands using several `.then()` statements: *SOM Browser* performs asynchronous operations using the `Promise` feature of ECMAScript 2015 <sup>2</sup>.

The crucial parts of the feature extraction code can be found in Listing 6. Since the SOM implementation in `dev/som-browser/src/background/calculateSOM.js` is logically identical to what was previously discussed, we refer to Section 3.1.2 rather than dissecting the contents of `calculateSOM()` separately.

<sup>1</sup><https://github.com/jonasmargraf/som-browser>

<sup>2</sup>[https://developer.mozilla.org/en-US/docs/Web/JavaScript/Reference/Global\\_Objects/Promise](https://developer.mozilla.org/en-US/docs/Web/JavaScript/Reference/Global_Objects/Promise)

```

284     if (this.state.files.find(e => !e.features)) {
285       console.log("Processing files...")
286       processFiles(this.state.files)
287       .then(files => this.setState({ files: files, loading: false }))
288       .then(() => {
289         console.log("Building map...")
290         createSOM(this.state.files, this.state.settings)
291         .then(som => {
292           this.setState({ som: som })
293           console.log(this.state)

```

Listing 5: `dev/som-browser/src/components/App.js: handleAnalyzeClick()` [excerpt]

```

66     // Framewise loop over audio and extract features
67     for (let start = 0; start < zeroPaddedSignal.length; start += bufferSize) {
68
69       let signalFrame = zeroPaddedSignal.slice(start, start + bufferSize)
70       let frameFeatures = Meyda.extract(featureList, signalFrame)
71       // we only use total loudness, not per band
72       frameFeatures.loudness = frameFeatures.loudness.total
73
74       // Append this frame's features to array of feature frames
75       for (let feature in frameFeatures) {
76         // Only use frames that have RMS > -60dBFS
77         (frameFeatures.rms >= 0.001) && features[feature].push(frameFeatures[feature])
78       }
79     }
80
81     // Get feature average
82     for (let feature in features) {
83       features[feature] = math.mean(features[feature])
84     }
85
86     // Add file duration to features
87     features.duration = decodedAudio.duration
88
89     // Pass averaged features to parent file
90     file.features = features
91     resolve(file)

```

Listing 6: `dev/som-browser/src/background/extractFeatures.js`: Audio features for each sound are extracted in `extractFeatures()` using the *Meyda* library (shown here is an excerpt). Formal definitions for each feature can be found in the equations of Section 2.1.

### 3.2.5 User Interface Components

The following paragraphs give an overview of the different React components that make up the GUI of *SOM Browser* (refer to Figure 4 for a screenshot of the program). An outline of the entire application interface is shown in Listing 7, which contains most of the `render()` function of `dev/som-browser/src/components/App.js`, including all components, their properties and functions.

**MenuBar** is the horizontal element across the top of the application window. It holds five buttons to perform audio analysis, load and export samples and store and recall calculated maps.

```

405     <div className="TitleBar"> <p>SOM Browser</p> </div>
406
407     <MenuBar
408       files={files}
409       onChange={this.handleFileListChange}
410       onFileClick={this.handleFileClick}
411       onAnalyzeClick={this.handleAnalyzeClick}
412       onSaveClick={this.handleSaveClick}
413       onLoadClick={this.handleLoadClick}
414       onExportClick={this.handleExportClick}
415       onPrintState={this.handlePrintStateClick}
416     />
417
418     <div className="leftPanel">
419       <input id="tab1" type="radio" name="tabs" defaultChecked/>
420       <label htmlFor="tab1">Sounds</label>
421       <input id="tab2" type="radio" name="tabs"/>
422       <label htmlFor="tab2">Settings</label>
423       <div className="content">
424         <div id="tabFileList">
425           <FileList
426             loading={this.state.loading}
427             files={files}
428             selectedFile={file}
429             onChange={this.handleFileListChange}
430             onFileClick={this.handleFileClick}
431             onAnalyzeClick={this.handleAnalyzeClick}
432             onSaveClick={this.handleSaveClick}
433             onLoadClick={this.handleLoadClick}
434           />
435         </div>
436         <div id="tabSettings">
437           <Settings
438             filesLength={files && files.length}
439             settings={this.state.settings}
440             onChangeSettings={this.handleChangeSettings}
441           />
442         </div>
443       </div>
444     </div>
445
446     <Map
447       som={this.state.som}
448       files={this.state.files}
449       progress={this.state.progress}
450       selectedFile={file}
451       onMapClick={this.handleMapClick}
452       onMouseLeave={this.handleMouseLeave}
453     />
454
455     <FileInfo file={file} />
456
457     <UserSelection
458       userSelection = {this.state.userSelection}
459       onClick={this.handleMapClick}
460       onUserSelectionUpdate={this.handleUserSelectionUpdate}
461     />

```

Listing 7: dev/som-browser/src/components/App.js: GUI Components

**FileList** is one of two options for content to display in the left panel. It shows a list of all loaded audio files. Each list element is rendered by a nested **FileListItem** component. When clicking an item in the list, the corresponding map square will be highlighted and the **FileInfo** panel on the right will display information about the file.

**Settings** is the other display option for the left panel. Here, a number of SOM parameters can be set, similar to the Max messages listed in Table 1.

**Map** is the main focal point of the application. It displays a grid of grey and white squares, where each white square represents one sound. The **Map** component is made up of three nested components: **MapNode** (the grey background squares that are rendered first), **MapSubNode** (the white squares representing sounds) and **MapLabel** (the label displaying the name of the sound over which the user is currently hovering).

**FileInfo** is the panel on the right side of the application. It displays details about the currently selected file, including its path, duration and audio feature values.

**UserSelection** is the horizontal element across the bottom of the application. It consists of eight instances of the nested component **UserSelectionSlot**, each representing a spot into which the user can drag a sound to keep as a reference or for later exporting.

### 3.2.6 Algorithm Extension: Forced Node Population

After the SOM algorithm was first implemented in *SOM Browser*, it became apparent that large parts of almost all created maps remained empty, which proved frustrating to interact with since these empty areas are essentially “dead” spots where no sounds are located and no interaction is possible. A method to counteract this phenomenon was deemed necessary.

In order to avoid “empty” nodes on the SOM — meaning nodes to which no input vectors are mapped — a post-processing extension was added to the original algorithm that inverts the mapping process, explicitly iterates over empty nodes and assigns each one that input vector which is closest. This algorithm extension, which we call Forced Node Population (FNP), executes the following sequence after the regular SOM has been calculated (refer to Listing 8 to see the actual source code implementation):

1. Select a random empty node.
2. Find the closest vector for that node and assign it to this node.
3. Remove this vector from the possible choices.
4. Repeat.

This process is only performed once for all nodes that are empty immediately after the initial SOM calculation. It is possible that the forced node population creates new empty nodes, but in order to minimize distortion introduced by this procedure (see Section 5.1), it is not repeated. The effect of the FNP extension can be clearly seen in Figure 7. For a closer look at the results of this algorithm extension, please refer to Section 5.1.5.

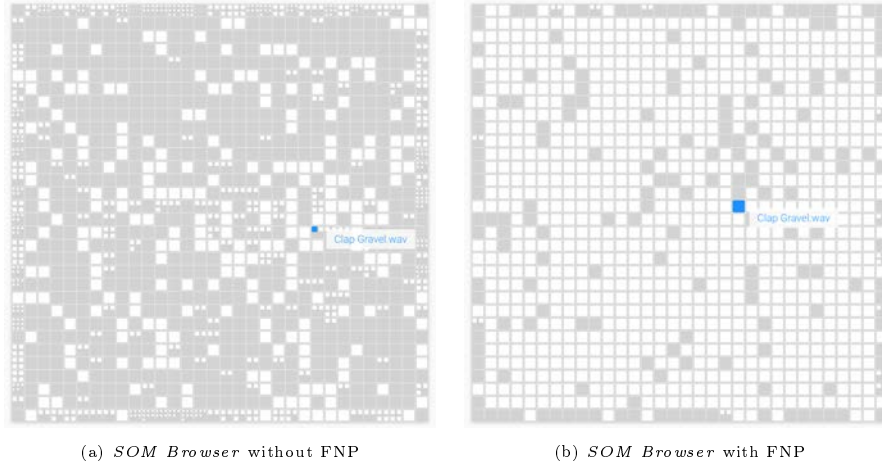


Figure 7: Comparison of *SOM Browser* maps with (7b) and without (7a) FNP. White squares each represent a sound, while grey squares denote unpopulated (“empty”) nodes. Smaller squares are used if multiple sounds are assigned to the same node on the map.

```

358 function populateEmptyNeurons(som) {
359
360   let emptyNeuronIndices = som.neuronAssignedFiles.map((e,i) => {
361     return (e === null ? i : false)
362   })
363   .filter(e => e !== false)
364
365   let tempVectors = som.normalizedData
366   let tempVectorIndices = tempVectors.map((e,i) => i)
367
368   while (emptyNeuronIndices.length >= 1) {
369     // Get random empty neuron, then remove from possible choices
370     let emptyNeuronIndex = math.pickRandom(emptyNeuronIndices)
371     emptyNeuronIndices.splice(emptyNeuronIndices.indexOf(emptyNeuronIndex), 1)
372     let emptyNeuron = som.neurons[emptyNeuronIndex]
373
374     let distancesFromNeuron = tempVectorIndices.map((e,i) => {
375       return math.norm(math.subtract(emptyNeuron, tempVectors[e]))
376     })
377
378     let nearestVectorIndex = tempVectorIndices[distancesFromNeuron.indexOf(
379       math.min(distancesFromNeuron))]
380     // Remove the found closest vector from its previously assigned neuron
381     // and instead assign it to the empty neuron.
382     let oldAssignedNeuronIndex = som.neuronAssignedFiles.findIndex(e =>
383       Array.isArray(e) && e.some(el => el === nearestVectorIndex))
384     som.neuronAssignedFiles[oldAssignedNeuronIndex].splice(
385       som.neuronAssignedFiles[oldAssignedNeuronIndex].findIndex(
386         e => e === nearestVectorIndex), 1)
387     som.neuronAssignedFiles[emptyNeuronIndex] = [nearestVectorIndex]
388
389     tempVectorIndices.splice(distancesFromNeuron.indexOf(
390       math.min(distancesFromNeuron)), 1)
391   }
392
393   return som
394 }

```

Listing 8: `dev/som-browser/src/background/calculateSOM.js`: The SOM algorithm extension FNP (see Section 3.2.6) is implemented in `populateEmptyNeurons()`.

## 4 Evaluation

In order to evaluate the SOM algorithm as implemented in this thesis, as well as the developed *Som Browser* application as a whole, a two-part process was employed. First, a set of numerical metrics was selected to quantify aspects of the algorithm we deem salient when judging its effectiveness for sound corpus organization. Second, a series of five semi-structured interviews was designed, conducted and qualitatively analyzed. The following sections go into detail about the selection of a data set of sound files for the evaluation, the metrics employed and the design of the interview.

### 4.1 Sound Corpus Selection

A crucial aspect for the evaluation of the work presented in this thesis is the choice of an appropriate data set of audio files to serve as a prototypical sound corpus. Ideally, two key conditions should be met by this corpus. It should be *ecologically valid*, meaning here that it should approximate a real-world sample library that would actually be used by contemporary music producers, and it should be a *well-established* data set which has been validated through use in other research, allowing for direct comparisons between results. Preferably, something akin to the Giant Steps data sets (Knees et al., 2015) for tempo and key detection should be used. In addition to identifying the aforementioned two conditions, the decision was made to only select “one-shot” drum and percussion sounds (meaning single instrument hits, no loops or other longer sounds) in order to evaluate a single, concrete use case and limit the scope of this evaluation.

Data sets used in previous research vary and it is often not possible to clearly establish provenance due to insufficient information being given by the authors (see for example Fried et al. (2014) and Shier et al. (2017), two papers which present important related work but fail to clearly identify the source of their employed sound files). Two established databases that have been cited in the literature are ENST-Drums (Gillet and Richard, 2006) and the RWC Music Database (Goto et al., 2002). However, neither of these data sets proved appropriate for this evaluation since they both contain only acoustic source material and, especially in the case of RWC, largely consist of longer musical passages instead of the single “one-shot” hits mentioned above.

For the reasons outlined above, we decided to forego the notion that the selected sound corpus be a data set well-established through previous research. Because of this, more emphasis is placed on the requirement for ecological validity. In order to maximize real-world conditions, the sample library *Drum Essentials* (Ableton AG, 2019b) was selected to serve as a sound corpus for this evaluation. It is a collection of samples created by the German music software company Ableton AG that is distributed to owners of the company’s flagship product, the DAW *Ableton Live* (Ableton AG, 2019a). As part of a commercially available product, this corpus of sound files does not just approximate a real-world sample library, it is an actual example of such a library and is, for the purpose of this thesis, considered representative of sample libraries used in a modern music production workflow. One additional benefit of using the selected sample library is the advantage of a single, clearly identifiable source of the data — it is made available as a professional product by Ableton. An alternative approach would have been to manually select sounds from places like *Freesound.org* (Font et al., 2013), where all files are licensed in a way



that makes them free to use, but their quality is not guaranteed to be consistent, or to scour through sample libraries shared on various online forums, which brings along issues of copyright and expired links, making it hard to trace the files’ origins.

The *Drum Essentials* collection as distributed by Ableton consists of 1181 one-shot samples, each in a separate audio file, as well as supplementary content, such as MIDI clips and effects presets. Only the raw audio files are used in the presented work. These sound files present a mixture of acoustic and electronic sounds stemming from a variety of drums and percussion instruments. The library is organized by instrument group, of which there are 17 in total. The names of these groups, as well as the number of sounds per group can be found in table 2. Some sound files appear in more than one group. These duplicates have been removed, so that every sound only appears once throughout the entire data set. The remaining number of sound files is 1081.

Drum Essentials	
Instrument Category	Count
Bell	19
Bongo	6
Clap	71
Conga	27
Cymbal	54
Electronic Percussion	49
FX Hit	64
Hihat	167
Kick	166
Misc. Percussion	64
Ride	40
Rim	65
Shaker	39
Snare	181
Tambourine	23
Tom	138
Wood	8

Table 2: Shown are sound file counts per instrument category of the *Drum Essentials* sample library.

## 4.2 Metrics for SOM Analysis

The optimal SOM for a given sample library should represent the input data with minimal distortion. Samples should be mapped evenly to the nodes and all areas of the map should be populated to maximize the usefulness of the interface space used.

In order to evaluate the SOMs created using the *SOM Browser* application and the *Drum Essentials* test data set, three core metrics are used: quantization induced by the SOM, map emptiness, and the ratio between nodes and their assigned vectors. These metrics and the motivation behind them are outlined further in the following paragraphs.

#### 4.2.1 SOM-Induced Quantization

Fundamental to the SOM principle is the idea of mapping vectors to their corresponding BMUs, those nodes that are closest to them (see section 2.2). Several vectors can be assigned to one node — this can also be thought of as a quantization process, where the magnitude of the difference between the positions of vector  $x_t$  and node  $m_c$  is the quantization error  $\Delta_t$  for that vector:

$$\Delta_t = ||x_t - m_c||. \quad (21)$$

To analyze the SOM, the distribution of quantization errors can be examined. Of special interest as a single value metric for map quality is the median vector quantization error:

$$\tilde{\Delta} = \text{median} \{ \Delta_1, \dots, \Delta_t \}. \quad (22)$$

In order to maximize information preservation, quantization errors should be minimized.

#### 4.2.2 Vector-Node Count

A second metric that was devised in order to quantify SOM quality is the count  $C_i$  of vectors  $x_1, \dots, x_t$  mapped to a node  $m_i$  and subsequently the distribution of those counts across the map. Ideally, this distribution should look like a single, narrow spike — meaning that (almost) all nodes have about the same number of vectors assigned to them, resulting in an even distribution of sounds across the *SOM Browser* map interface. SOM parameters should be chosen to approximate a uniform vector-node count for all nodes. For our purposes, we examine the mean vector-node count of all  $n$  nodes for a given map:

$$\overline{C} = \frac{1}{n} \sum_{i=1}^n C_i \quad (23)$$

#### 4.2.3 Map Emptiness

Another relevant aspect of the created SOMs, and the third metric employed here, is how much of the map remains “empty”, meaning how many nodes were not assigned any vectors. We define this “map emptiness” metric  $ME$  as the number of nodes  $m_1, \dots, m_n$  whose vector-node count  $C_i = 0$  (see section 4.2.2), divided by the total number  $n$  of nodes  $m_i$ :

$$ME = \frac{1}{n} \sum_{i=1}^n \delta_{C_i 0}, \quad (24)$$

where

$$\delta_{C_i 0} = \begin{cases} 0, & \text{if } C_i \neq 0, \\ 1, & \text{if } C_i = 0. \end{cases} \quad (25)$$

For the purpose of making optimal use of the space allotted to the map in the *SOM Browser* GUI, emptiness should be minimized so that users encounter the least amount of “blind spots” possible.

#### 4.2.4 Influence of Forced Node Population

Since the concept of Forced Node Population is an addition to the SOM algorithm introduced in this work (see section 3.2.6), its influence on the SOM should also be evaluated. Therefore, the aforementioned metrics were calculated both with and without FNP.

### 4.3 Semi-structured User Interviews

In order to evaluate the *SOM Browser* application prototype presented in this thesis, five semi-structured interviews with working audio professionals were conducted. These interviews were conducted by the author and consisted of a set of questions as well as observed user interaction with the prototype software. For this evaluation, a guide including questions outlining the structure of the interview as well as a set of ratings scales was created. Subjects were asked about their experience with sample libraries and their current workflow, and to interact with a sample library in a file browser environment as well as using the *SOM Browser* software. Audio from the conversations was recorded and subsequently qualitatively analyzed.

#### 4.3.1 Motivation to Conduct Interviews

This evaluation procedure entails two aspects, namely a semi-structured interview series and a qualitative analysis of the collected responses. The decision to conduct qualitative interviews stems from the exploratory nature of the presented work. In order to assess the merit of the developed interface in its present state, direct feedback from potential users was sought, which Lazar et al. (2017) refers to as “fundamental to human-computer-interaction (HCI) research” (see Lazar et al. (2017, p.187)). But the motivation for a direct conversation with users was not only to evaluate the presented interface proposition, but also for these interviews to serve as an exploration of users’ current situation, to hear about their own experience of it and to see what advantages and shortcomings they identify in their present workflows. In short, these interviews were motivated by a desire to gain some understanding of the complex situation that is sample library interaction in a music production environment and to gauge initial reactions to the developed prototype alternative. The semi-structured approach was chosen in order to be able to react to interviewees’ responses more freely and allow the interviewer to ask follow up questions when deemed necessary. Naturally then, the gathered responses cannot simply be quantified, which makes a qualitative approach to their analysis a fitting choice.

There are of course downsides to the chosen approach. Conducting interviews is time-consuming, as it has to be done on a one-on-one basis and often (as in the case of this work) in person. After the interview is over, additional time and effort goes into transcribing and annotating the responses. This severely limits the number of participants that can feasibly be recruited for a study, as is evident by the small number of five participants here. Lazar et al. (2017) identifies another disadvantage of interviews: “[...] data collection that is separated from the task and context under consideration [...] suffer[s] from problems

of recall. [...] [I]t is, by definition, one step removed from reality” (Lazar et al., 2017, p.188ff.). Because of this, we follow the authors’ suggestion of combining the interview with user observation.

#### 4.3.2 Interview Subject Selection

The SOM Browser application is not aimed at the general population. Instead, it has been designed for specialized users that work in modern music production, as they constitute the potential future user base of an application like the one presented here.

In order to increase the validity and relevance of potential subjects’ responses, the decision was made to interview only working professionals for this evaluation and to not include hobbyists or people without any experience in music production.

Subjects were recruited by inquiring about qualified candidates (in other words, people working professionally in modern music production) in the wider circle of acquaintances of the author. No compensation was offered and only sparse information about the nature of the research was given beforehand in order to minimize the possibility of instilling biases in subjects. Most importantly, subjects were asked to participate in an interview about sample library organization, but were not told that they would be shown software developed by the author.

#### 4.3.3 Informed Consent Form

For the purpose of documenting participants agreement to be interviewed, an informed consent form was created for the interview series. This document outlines basic information about the purpose and content of the interview and its duration. It also lists all data that will be collected and explains the procedure used for data anonymization in order to protect subjects’ privacy. Lastly, it informs participants of their rights to withdraw their consent to the usage of their data for research purposes and have it erased. This form was based on a template provided by the ethics board of Technische Universität Berlin (TU Berlin) on their website (TU Berlin, 2019). The form used by the author is included with the digital submission of this thesis at the location `evaluation/Eval_ConsentForm.pdf`.

#### 4.3.4 Test Subject Code Design

To ensure proper data anonymization, a test subject code was used. This code is comprised of a series of letters and numbers and was created at the beginning of the interview by the subjects themselves according to a set of instructions. All data and responses of the subjects were directly labelled with this code, so that individuals’ names were never used. This code design procedure was again based on a template by the ethics board of TU Berlin and can be found on the same website as the information concerning consent forms (TU Berlin, 2019). The instruction sheet that was distributed to subjects can be found as part of the digital submission of this thesis in `evaluation/Eval_TestSubjectCode.pdf`

#### 4.3.5 Interview Structure

The guide developed for this interview is included with the digital submission of this thesis in `evaluation/Eval_Questionnaire.pdf`. It outlines a three part structure: first, some

general questions about subjects’ usage of sample libraries are posed. Second, some guided interaction with a predetermined sample library in a traditional file browser structure on a computer follows. In the third section, the SOM Browser application is finally introduced and subjects are asked to use it and describe their impression of it.

#### 4.3.6 Question Design

The general composition employed for most questions is twofold, combining closed- and open-ended approaches: first, participants are asked to give a rating on a predefined scale (see 4.3.7 below). Then, participants are free to elaborate on their answer and explain their rating. If they don’t initiate this themselves, a follow-up question along the lines of “Could you tell me why you chose this rating?” is asked.

#### 4.3.7 Selection of Ratings Scales

In order to record subjects’ ratings, 6 point Likert scales were used (as is common in Human-Computer Interaction (HCI) research, see Lazar et al. (2017, p.31, p.93)). The difference between even and uneven anchor counts in Likert scales lies in the presence (in the case of uneven anchor counts) or lack (for even counts) of a “neutral” middle option. Choosing scales without neutral mid-points was motivated by a desire to encourage subjects to make a definite choice with regard to their rating. For a short look at the effects of eliminating the mid-point, see Garland (1991). The scales presented to subjects were explicitly labeled textually instead of numerically. The anchor points were designed using two polar adjectives (such as “positive” and “negative”) and a consistent, three-tiered set of adjective qualification with “very” marking the strongest option, followed by the adjective without qualifier and then “somewhat” as the weakest variant. The resulting scale for a positive/negative rating is composed of the following anchors: very positive, positive, somewhat positive, somewhat negative, negative, very negative. The selection of these qualifiers and appropriate anchors in general was inspired partially by Vagias (2006). The full set of scales used for the conducted interviews can be found in `evaluation/Eval_Questionnaire_LikertScales.pdf`.

#### 4.3.8 Questions Used

In section 1, which serves as an introduction for the interviewee, general administrative requirements such as the signing of the consent form and a topical introduction of the research are taken care of. This is then followed by two simple Yes/No questions to establish whether the subject works with third-party and/or personally created sample libraries (see questions 1.1 and 1.2).

Section 2 begins with a presentation of the *Drum Essentials* sample library to the subject. This presentation includes the information that it is a library of drum samples that consists of around 1000 sound files which are organized in subfolders according to the respective instrument, such as kick drum, snare drum, hi-hat, and so forth. The interviewee is invited to explore the sample library using the laptop that it is being presented on.

Then, in question 2.1, subjects are asked to describe how to approach familiarizing themselves with the provided sample library in order to use its contents in a hypothetical work project of theirs.

Question 2.2 follows this up with a request for a rating of the subject's level of satisfaction with the workflow that they outlined.

In the third and final section of the interview, the SOM Browser software is introduced to participants. At first, a general overview of the interface is given, in which the interviewer mentions the map layout in the middle (without explaining the nature of its organization), the file list on the left, the file info panel on the right and the favorites bar at the bottom.

The subject is then asked to try out the software and explore its interface for a short period of time. Thereafter, they are asked to give a rating of their overall first impression of the software on a positive/negative scale (see question 3.1). Then, a follow-up question about their opinion on what does or does not work is posed.

In question 3.2, subjects are required to rate the interface's ease of use.

Question 3.3 inquires specifically about the understandability of the language used.

3.4 and 3.5 are open-ended questions aimed at subjects' interpretation of the organization of sounds in the map layout: 3.4. asks what subjects think about the organization, while 3.5 inquires specifically about a guess as to what the axes represent.

Question 3.6 then asks subjects to state whether or not they have a preference between the traditional file browser layout presented in section 2 or the *SOM Browser* interface shown in section 3.

The last ratings question of the interview, 3.7 requests interviewees to assess their level of comfortability with the software.

Finally, in 3.8 subjects are asked if they would consider using the presented software tool and what changes they would like to see.

The full interview guide including all questions can be found as part of the digital submission of this thesis at the location `eval/Eval_All_Documents.pdf`.

## 5 Results

This chapter contains the results of our search for an optimal SOM for the *Drum Essentials* sample library and the findings of the conducted user interviews.

### 5.1 SOM Metrics

*SOM Browser* exposes several variables of the algorithm, offering the user influence over the resulting SOM. The following sections show the effects of these different user-controllable parameters on the metrics outlined in Section 4.2. We then compare the three different learning rate factor approaches that were implemented, including a differentiation between different magnification control values for the BDH method (see Section 2.2.4). Finally, an ideal set of parameters for the selected sound corpus is chosen and the resulting map is presented. This map is then also calculated using our FNP algorithm extension introduced in 3.2.6 and the results compared.

Ideally, we'd like the map shown in *SOM Browser* to have one or two vectors per node (meaning each map square only represents one or two sounds) and no empty nodes. We therefore chose a map size of  $32 \times 32 = 1024$  as it is the closest square number to 1081, the number of samples contained in our chosen corpus.

#### 5.1.1 Effects of Training Duration

The influence of training duration on the quality of the produced map can be seen in Figure 8. As duration increases, all metrics improve. As can be seen in the boxplot diagram, the distribution of quantization errors narrows in on the median with longer training as well.

A practical consideration to be noted is that training duration directly impacts processing time. For our measurements, which were run on a fairly recent consumer laptop, training for  $10^4$  steps took on the order of minutes, while training for  $10^5$  steps finished in about 30 minutes, and the training for  $10^6$  steps had to be run over night, as it took several hours. Because of these practical considerations, which we assume to reflect the needs of potential users of the software, all following maps were trained for  $10^4$  steps.

#### 5.1.2 Effects of $\alpha_{initial}$

The effect of the choice of  $\alpha_{initial}$ , the starting value for the “gain control” of the node adjustment in response to each vector during training, is shown in Figure 9. Higher  $\alpha_{initial}$  values yield slightly better results on quantization errors, but don't improve vector node counts or map emptiness. Although this effect does not appear to be dramatic, it is in line with Kohonen (1990), who suggests a starting value of  $\alpha = 0.9$ . We therefore choose  $\alpha_{initial} = 0.9$  for the next maps.

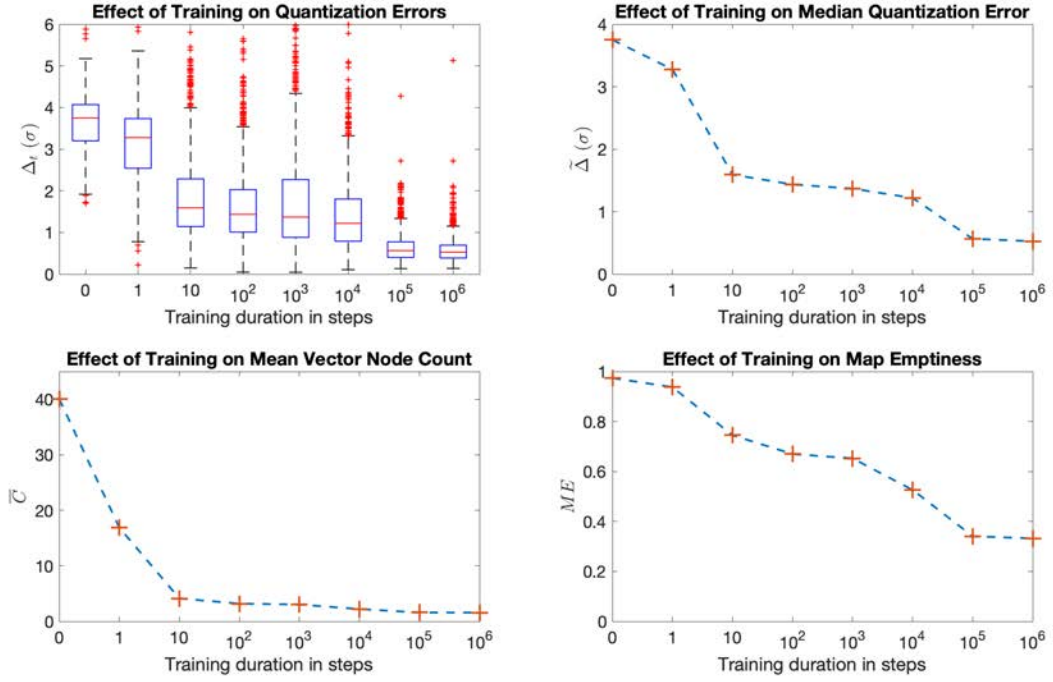


Figure 8: *SOM Browser*: effect of training duration on SOM metrics. SOM parameters: `mapSize` =  $32 \times 32$ , linearly decreasing  $\alpha$  with  $\alpha_{initial} = 0.9$ ,  $r_{start} = 6.4$ ,  $r_{end} = 1$

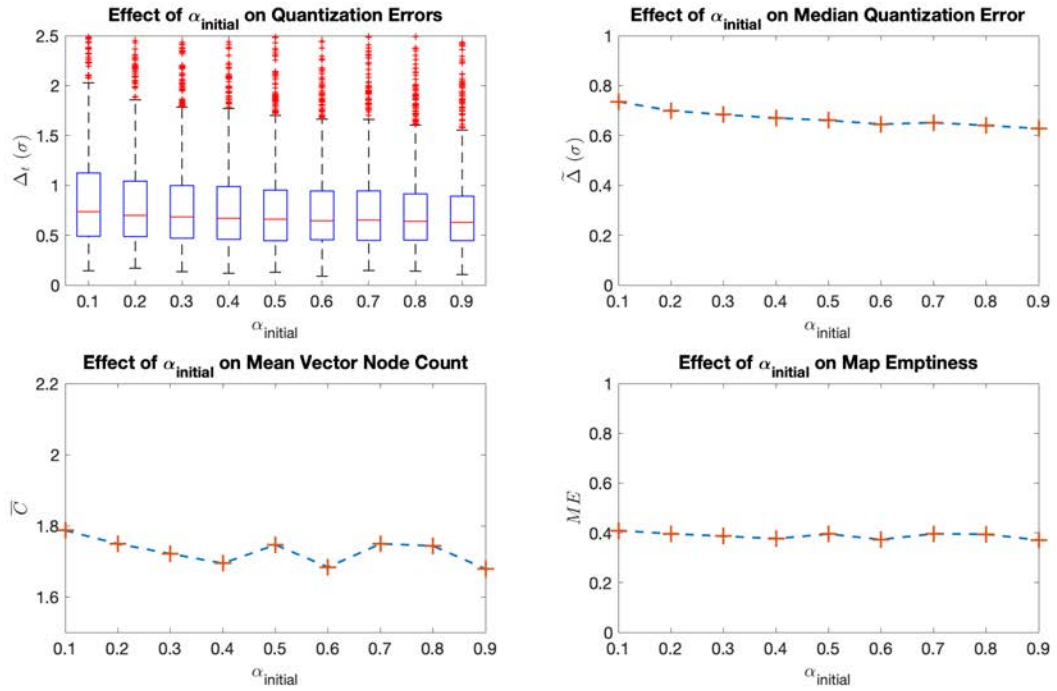


Figure 9: *SOM Browser*: effect of  $\alpha_{initial}$  on SOM metrics. SOM parameters: `mapSize` =  $32 \times 32$ , training duration =  $10^4$  steps, linearly decreasing  $\alpha$  with  $\alpha_{initial} = [0.9, \dots, 0.1]$ ,  $r_{start} = 6.4$ ,  $r_{end} = 1$



### 5.1.3 Effects of Radius Size

Next, we examine the effects of the neighborhood function radius (Figures 10 and 11). This kernel function, which controls how far around the best matching unit nodes will be adjusted during each training step, decreases from a starting value  $r_{start}$  to a final value  $r_{end}$ . Kohonen (1990) recommends starting with a rather large radius and ending with a value of one unit. This recommendation does not coincide with our measurements, which obtained best results when using a narrow starting kernel radius of  $r_{start} = 2$  that decreases to  $r_{end} = 10^{-1}$ . It should be noted that although we achieved the best results with  $r_{start} = 2$ , we used  $r_{start} = 16$  for the comparison of  $r_{end}$  values in order to avoid a kernel that increases in size as training progresses.

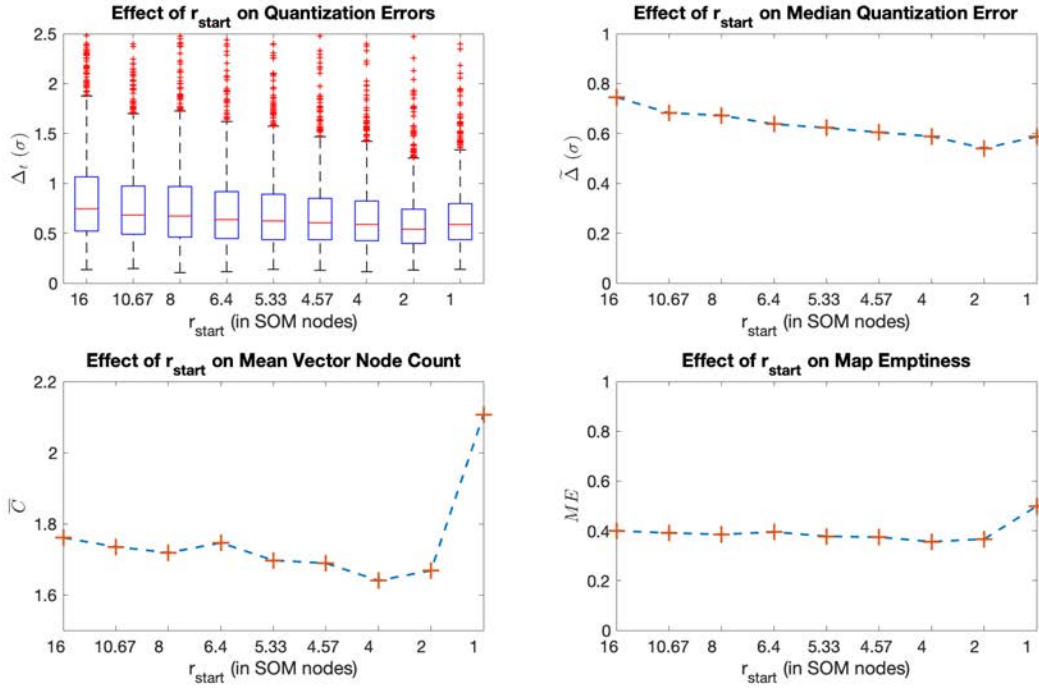


Figure 10: *SOM Browser*: effect of  $r_{start}$  on SOM metrics. SOM parameters: `mapSize` =  $32 \times 32$ , training duration =  $10^4$  steps, linearly decreasing  $\alpha$  with  $\alpha_{initial} = 0.9$ ,  $r_{start} = [16, \dots, 1]$ ,  $r_{end} = 1$

### 5.1.4 Learning Rate Type Comparison

There are three different approaches to the learning rate factor  $\alpha$  that are implemented in *SOM Browser*: linearly decreasing, reciprocally decreasing (“Inverse”), and the local adaptive learning rate as proposed by Bauer et al. (1996) (“BDH”). Differences between these are shown in Figure 12 and Table 3. While “Inverse” shows worse results than “Linear”, improvements can be seen when using BDH.

BDH offers another parameter,  $m$ , that can be controlled. Bauer et al. (1996) suggest values of  $m = -1, -0.5$  or  $-0.25$ . We separately measured the effects of values for  $m$  of  $[1.0, 0.9, \dots, -0.9, -1.0]$ , the results for which can be seen in Figure 13. Interestingly, median quantization errors, mean vector node counts and map emptiness all reach a minimum at  $m = 0$ , which effectively results in a constant learning rate of  $\alpha = \alpha_{initial}$ , as

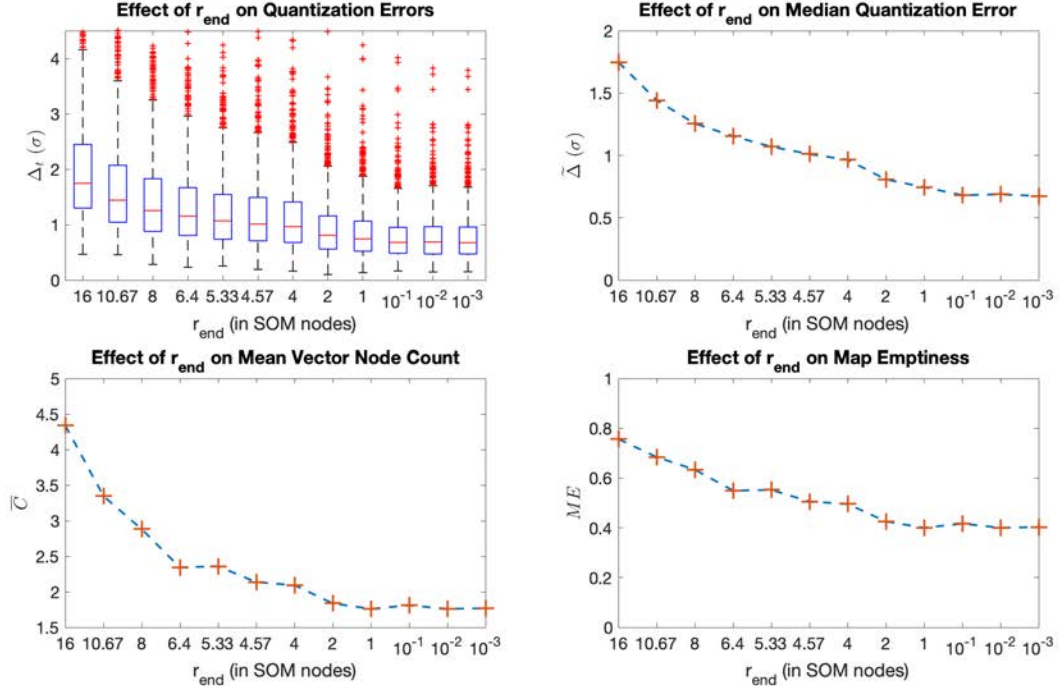


Figure 11: *SOM Browser*: effect of  $r_{\text{start}}$  on SOM metrics. SOM parameters: `mapSize` =  $32 \times 32$ , training duration =  $10^4$  steps, linearly decreasing  $\alpha$  with  $\alpha_{\text{initial}} = 0.9$ ,  $r_{\text{start}} = 16$ ,  $r_{\text{end}} = [16, \dots, 1]$

the adaptive local learning term of the BDH equation equals 1 when  $m = 0$  (see Section 2.2.4). At the same time, the distribution of errors increases for positive values of  $m$  and is generally larger than for  $m < 0$ . In order to strike a balance between a small median quantization error and a narrow error distribution,  $m = -0.2$  was chosen for the final map.

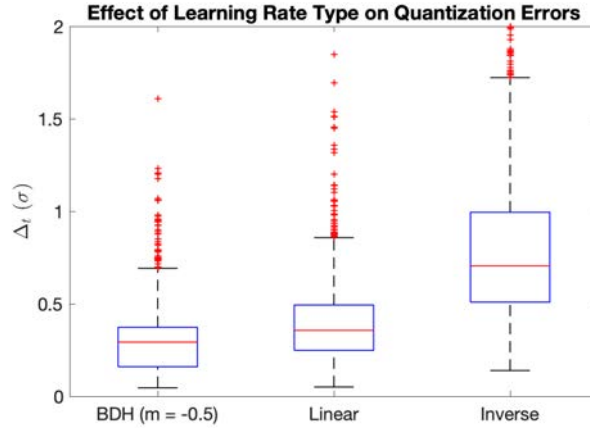
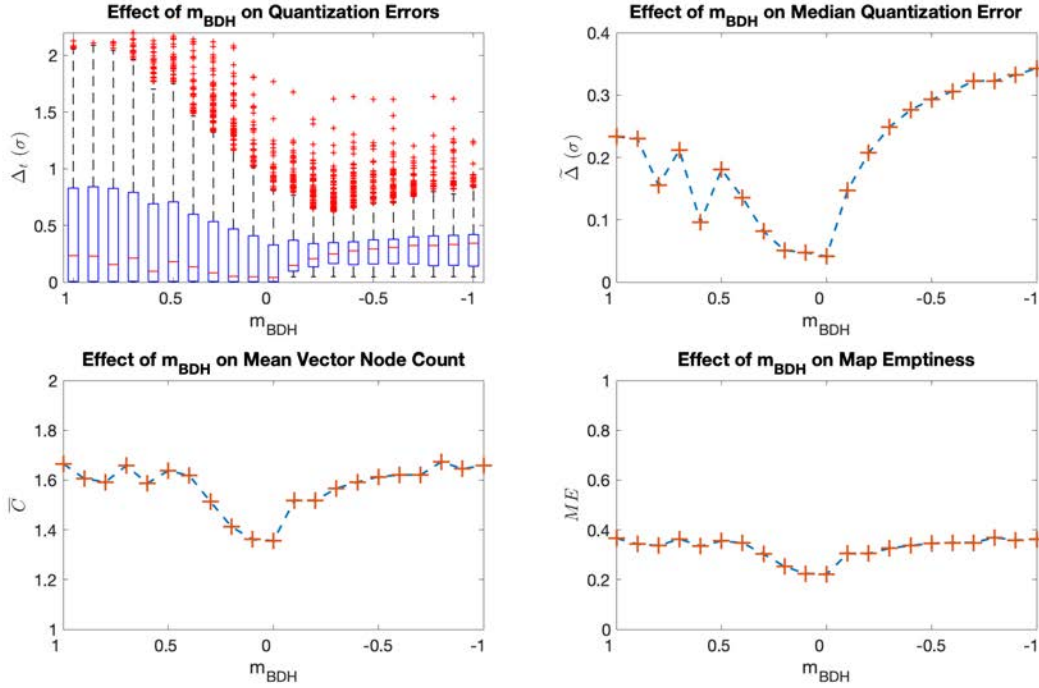


Figure 12: *SOM Browser*: comparison of linearly decreasing, reciprocally decreasing (“Inverse”) and BDH learning rate factors. SOM parameters: `mapSize` =  $32 \times 32$ , training duration =  $10^4$  steps,  $\alpha_{\text{initial}} = 0.9$ ,  $r_{\text{start}} = 2$ ,  $r_{\text{end}} = 0.1$

Comparison of SOM Learning Rate Types

	BDH	Linear	Inverse
Median Quantization Error $\tilde{\Delta}$	0.29	0.36	0.71
Mean Vector Node Count $\bar{C}$	1.61	1.49	3.16
Map Emptiness $ME$	0.35	0.29	0.67

Table 3: Comparison of SOM Learning Rate Types

Figure 13: *SOM Browser*: BDH: Effect of Magnification Control  $m$ . SOM parameters: `mapSize` =  $32 \times 32$ , training duration =  $10^4$  steps,  $\alpha_{initial} = 0.9$ ,  $r_{start} = 2$ ,  $r_{end} = 0.1$ 

### 5.1.5 Final Map and Influence of FNP

Based on the results presented above, a final map for the *Drum Essentials* sound corpus was created with the parameters `mapSize` =  $32 \times 32$ , training duration =  $10^4$  steps,  $\alpha_{initial} = 0.9$ , BDH learning rate factor,  $m = -0.2$ ,  $r_{start} = 2$ , and  $r_{end} = 0.1$ . A second map was calculated using the same parameters, but also using the FNP algorithm extension we developed. Comparisons of the map with and without FNP are shown in Table 4 and Figures 14 and 15. It can be seen that while FNP introduces more distortion into the map (indicated by higher quantization errors), it does improve both mean vector node counts and map emptiness.

### Influence of Forced Node Population (FNP)

	With FNP	Without FNP
Median Quantization Error $\tilde{\Delta}$	0.25	0.20
Mean Vector Node Count $\bar{C}$	1.29	1.45
Map Emptiness $ME$	0.18	0.27

Table 4: Influence of FNP

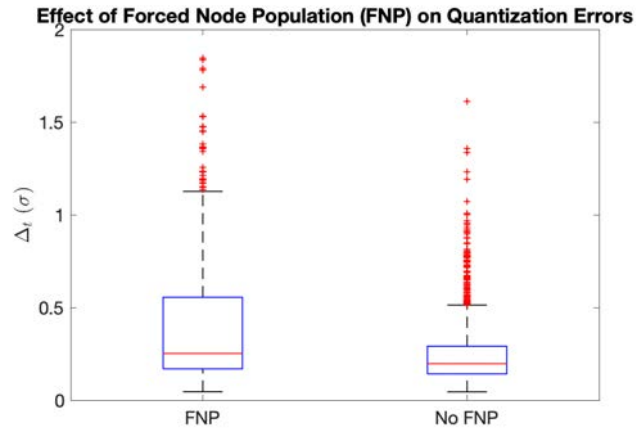


Figure 14: *SOM Browser*: Comparison of FNP influence on SOM. SOM parameters: `mapSize` =  $32 \times 32$ , training duration =  $10^4$  steps,  $\alpha_{initial}$  = 0.9, BDH learning rate factor,  $m = -0.2$ ,  $r_{start} = 2$ ,  $r_{end} = 0.1$

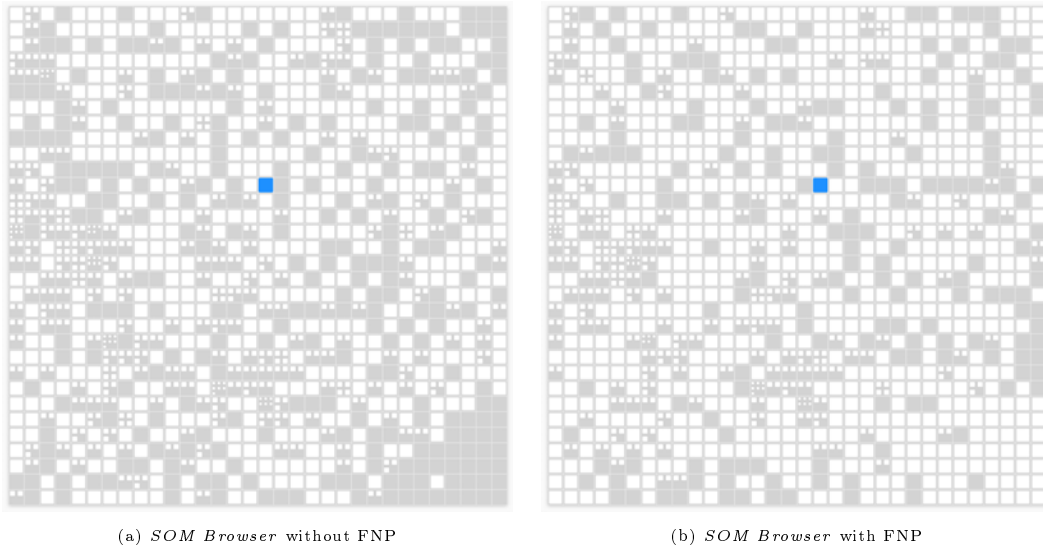


Figure 15: *SOM Browser*: Visual Influence of FNP on Map of the *Drum Essentials* sound corpus. The blue mark denotes the same file in both maps. SOM parameters: `mapSize` =  $32 \times 32$ , training duration =  $10^4$  steps,  $\alpha_{initial}$  = 0.9, BDH learning rate factor,  $m = -0.2$ ,  $r_{start} = 2$ ,  $r_{end} = 0.1$

## 5.2 Interview Results

Following the presentation of an optimal set of parameters for *SOM Browser* to display the *Drum Essentials* corpus, the rest of this chapter is dedicated to the results of the conducted user interviews, whose design is explained in Section 4.3.

The group of interview subjects consisted of one woman and four men with an average age of 36 years and an average 15 years of experience in the audio industry. Subjects described their profession as one or several of the following: composer, producer, DJ, performer, sound / mixing engineer, sound designer.

Subject responses to the ratings questions described in Section 4.3.8 can be found in Figure 16.

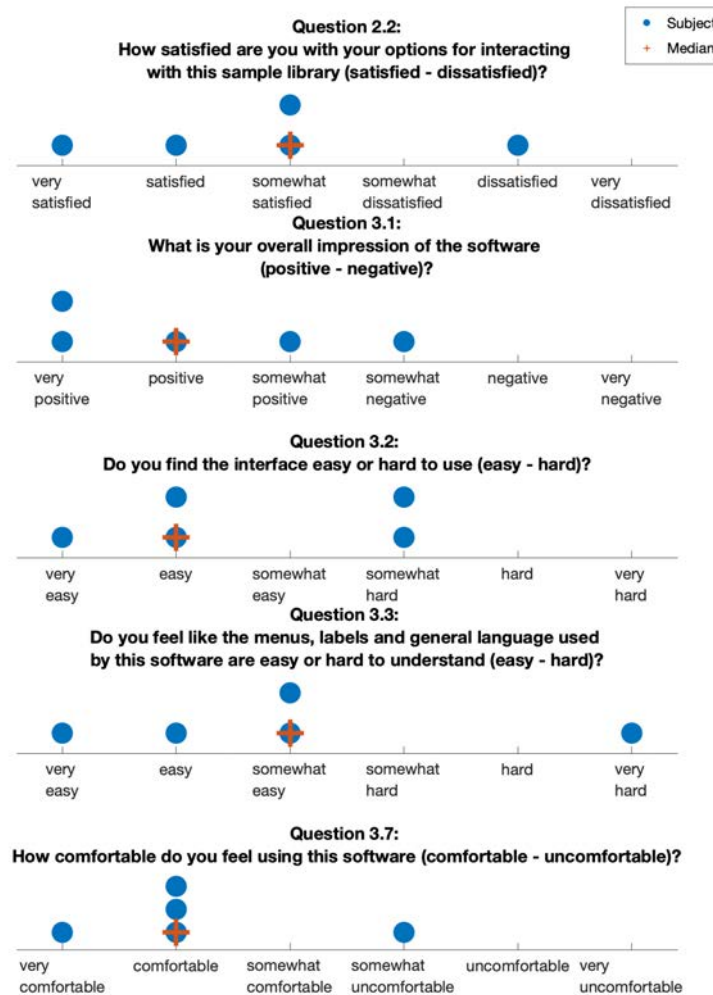


Figure 16: Likert scale ratings by interview subjects for questions concerning satisfaction with their current sample library workflow and first *SOM Browser* impressions

Recordings of the conducted interviews were transcribed by the author. These transcriptions were examined for instances relating to the interview questions and the relevant instances marked and subsequently coded using an *emergent coding* approach (Lazar et al., 2017, p.304). These codes are presented here using matrix data displays as proposed by Saldaña (2015, p.254) and Henwood and Pidgeon (2003).

### 5.2.1 Established Workflow Responses

Sections 1 and 2 of the interview questionnaire inquire about subjects' current workflow practices. Recorded responses are grouped into the categories *Description* (see Table 5) and *Assessment* (see Table 6). A flowchart of established workflow practices was extracted from these responses and is displayed in Figure 17.

Description of Established Sample Library Workflow

Code (Subject Count)	Example	Summary
<b>Mental Representation (3)</b>	"I know exactly what I'm looking for"	Subjects often have a clear <b>mental representation</b> of the sound they are searching.
<b>Goal Pursuit (3)</b>	[I listen to sounds] "[u]ntil I find the one that I want"	Subjects will only look for sounds until they find something that satisfies their immediate needs.
<b>Search Algorithm (5)</b>	"I would just go through every folder [...] and listen carefully to every sound"	Subjects describe three <b>search "algorithms"</b> : sequential (listening in alphabetical order), name-based (looking at file or subfolder names) and random search (arbitrarily selecting samples).
<b>Contextual Evaluation (3)</b>	"quickly go through the sounds while the track is playing and then find one that kind of fits"	The ability to audition sounds in the <b>context</b> of the relevant project is important.
<b>Iteration (4)</b>	"I will have like eight different kick drums [...] and then I go through them again as another iteration of choice."	Subjects will select a variety of samples as potential candidates and then perform another search among the selected subset.
<b>Frustration (5)</b>	"it takes a lot of time actually and it's not the most fun part"	Looking through lists of samples sequentially is perceived to cause <b>frustration</b> .

Table 5: Established sample library workflow as described by subjects. Shown are response codes along with example data and interpretive summary. The number shown in parentheses behind each code denotes how many individual subjects gave responses corresponding to that code.

### Assessment of Established Sample Library Workflow

Code (Subject Count)	Example	Summary
<b>Requires Organization (2)</b>	“if I was organized and I had my 5000 sounds from the past five years it [would] be really nice”	Subjects note that their current workflow relies on sample libraries that are organized in some way and note sources of <b>frustration</b> such as lost or duplicate files.
<b>Requires Experience (1)</b>	“experience [...] is probably the key”	<b>Experience</b> (both in a general professional sense and specific to the sample libraries at hand) is mentioned as a factor for an efficient, successful workflow.
<b>Good Enough (3)</b>	“It could be better but it's okay. Like, it works in most of cases.”	Current workflow practices are deemed “ <b>good enough</b> ”, but subjects are interested in alternative approaches.
<b>Time-Consuming (3)</b>	“[H]ow to listen to all this?”	Subjects remark upon the amount of time and effort that go into searching through sample libraries.
<b>Overwhelming (2)</b>	“it was overwhelming [...] to look through all this”	Finding relevant samples in a library is described as <b>overwhelming</b> .
<b>Alphabetic Bias (1)</b>	“I think it makes no sense that I'm mainly choosing from the first half of the alphabet”	Sample selection is influenced by alphabetical name ordering. Typically, samples positioned towards the beginning of an alphabetical list are more likely to be chosen.

Table 6: Subjects’ assessment of their established sample library workflow. Shown are response codes along with example data and interpretive summary. The number shown in parentheses behind each code denotes how many individual subjects gave responses corresponding to that code.

#### 5.2.2 SOM Browser Responses

The third and largest section of the questionnaire assesses subjects’ first impressions of the *SOM Browser* software. Responses varied between individual subjects, but can generally be grouped into *positive* and *negative* statements. These are presented in Tables 7 and 8. Notable *positive* responses were prompted by the **visual design** of the software, as well as the creative potential for using the software as an **instrument** because of the **gestural interaction** it facilitates. Looking at participants’ *negative* responses, the organization of the map interface was seen as **incomprehensible**, with subjects not able to directly discern an overarching order (although some areas of similarity were identified), which in turn led to some questioning of the app’s **usefulness** in its current state.

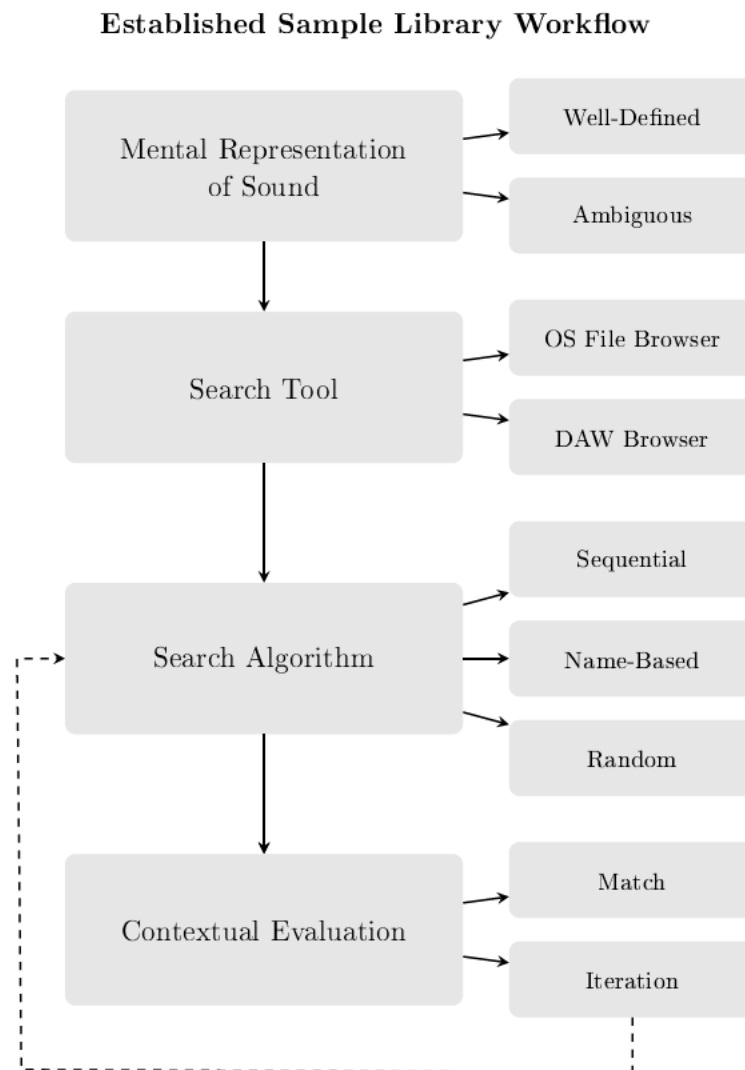


Figure 17: Flowchart of established sample library workflow as described by interview subjects



## SOM Browser: Positive Responses

Code (Subject Count)	Example	Summary
<b>Visual Design (4)</b>	“Visually, it makes sense.”	Subjects characterize the <b>visual design</b> of the software as appealing.
<b>Gestural Interaction (2)</b>	“for expressive gestures as a performance tool, it’s fantastic as a continuous thing”	Continuous playback with trackpad/mouse gestures (while holding down Shift) is seen positively when using <i>SOM Browser</i> more like an instrument.
<b>App as Instrument (2)</b>	“this is an instrument”	Subjects remark upon potential creative use of the software by not just using it to select samples, but also treating it as an <b>instrument</b> by itself.
<b>User-Friendly (3)</b>	“Even if I would be starting, like, it’s not confusing, [...] it’s clear”	Subjects describe use of the software as <b>user-friendly</b> .
<b>Favorites Selection (2)</b>	“I also find the favorites pretty good”	The <i>Favorites</i> bar at the bottom of <i>SOM Browser</i> can be seen as positive.

Table 7: Positive responses given by subjects after using *SOM Browser*. Shown are response codes along with example data and interpretive summary. The number shown in parentheses behind each code denotes how many individual subjects gave responses corresponding to that code.

## SOM Browser: Negative Responses

Code (Subject Count)	Example	Summary
<b>Incomprehensible Map Organization (5)</b>	“I don’t get the order, that’s frustrating.”	Subjects are not able to discern any logical order within the map.
<b>Usefulness (2)</b>	“right now it’s beautiful but it makes no sense unfortunately”	Subjects question the <b>usefulness</b> of the software in its current state.
<b>Confused by Empty Nodes (3)</b>	“Why are there a few grayed out?”	Empty nodes on the map confuse subjects.
<b>Overwhelming (1)</b>	“looking at these many, many tiny squares gives me anxiety”	The map interface is seen by some subjects as <b>overwhelming</b> .
<b>No File Labels (1)</b>	“here it’s just white bricks”	The fact that no files besides the current selection are labelled on the map is seen negatively by some subjects.
<b>Unnecessary Interface Elements (2)</b>	“Eliminate the top bar, eliminate all the bottom just keep this [points to map in center]”	Subjects question the need for interface elements besides the central map display.

Table 8: Negative responses given by subjects after using *SOM Browser*. Shown are response codes along with example data and interpretive summary. The number shown in parentheses behind each code denotes how many individual subjects gave responses corresponding to that code.

### 5.2.3 Workflow Comparison: SOM Browser vs. Established Workflow

Response codes concerning the comparison of *SOM Browser* and established workflows are shown in Table 9. Subjects stated a clear preference for their established workflows, but remarked upon the **potential** of the software, particularly if it could be presented as an **optional workflow** that integrates with existing production environments (something that is echoed in the feature requests in Section 5.2.4).

**Workflow Comparison: SOM Browser vs. Established Workflow**

Code (Subject Count)	Example	Summary
<b>Preference for Established Workflow (4)</b>	"I wouldn't want to miss my old way of looking for stuff."	Subjects state a preference for their established way of working with samples.
<b>Optional Workflow (3)</b>	"I would go for the traditional and just be presented with this like [an] alternative"	Subjects would like to incorporate <i>SOM Browser</i> into their workflow if it was well-integrated into their existing tools and could be used as an alternative view mode.
<b>Potential (4)</b>	"I could imagine, that this if you use it a bit and you get used to it, yeah, it could make things faster actually"	The potential for certain workflow improvements such as increased speed and removal of alphabetical bias is acknowledged.

Table 9: Subjects' responses when asked to compare *SOM Browser* to their established workflows and state a preference. Shown are response codes along with example data and interpretive summary. The number shown in parentheses behind each code denotes how many individual subjects gave responses corresponding to that code.

### 5.2.4 SOM Browser: Feature Requests

Lastly, subjects were asked what changes and additions they would like to see in the presented software. These responses can be found in Table 10. Two particularly noteworthy responses were the desire for **DAW integration** of the functionality of *SOM Browser* and the ability to navigate the map using **Arrow keys**.

## SOM Browser: Feature Requests

Code (Subject Count)	Example	Summary
<b>DAW Integration (2)</b>	“I would really like to see it as a plug-in or inserted in the production environment that I have. I would definitely not use it if it's a standalone thing.”	Subjects want to <b>integrate</b> the functionality of <i>SOM Browser</i> into their established production environment, either in plug-in form or directly within the DAW.
<b>Arrow Key Navigation (3)</b>	“You need the arrows.”	The need for granular map navigation using the keyboard's <b>Arrow keys</b> was mentioned repeatedly by subjects.
<b>User-Definable Map Organization (3)</b>	“be able to choose what each axis is”	The ability to more explicitly control how maps are organized.
<b>Pre-Categorization (2)</b>	“a pre-categorization would be nice”	Subjects express the wish to incorporate some sort of <b>categorization</b> to precede map calculation into the software.
<b>File Limit (2)</b>	“it's about limiting the squares”	Subjects would like to <b>limit</b> the total number of files that can be displayed on a map. By preventing larger maps, the interface would potentially become less overwhelming.
<b>Multi-Page Maps (1)</b>	“[...] something that's like a page based. Like, you go through the folders or you just type in quickly, just give here all the kicks or only all the snares”	Larger sample libraries could be spread out across several <b>pages</b> , potentially based on the previously mentioned <b>pre-categorization</b> .
<b>Feature Filters (2)</b>	“[If] I'm looking for a sample that [...] is very short, then you would have to be able to filter this somehow”	Incorporate the option to <b>filter</b> samples on the map based on certain feature values or ranges.
<b>Color Coding (3)</b>	“all the dark and bass heavy stuff [...] in another color than [...] all the tsh-tsh-tsh stuff [makes high frequency noises]”	Introduce <b>color</b> as an additional dimension to display information about the files on the map.
<b>Touchscreen Support (1)</b>	“Only if it's like on a touchscreen”	Support for <i>SOM Browser</i> on <b>touchscreen</b> devices.
<b>Color Customization (1)</b>	“maybe people can choose their colors”	Give users the option to <b>customize</b> interface <b>colors</b> .
<b>Sample Retriggering (1)</b>	“when I'm on a square I need a way to retrigger”	Ability to trigger the same sample repeatedly without clicking multiple times (presumably by pressing the Spacebar).
<b>Larger Font Size (2)</b>	“the letters [need] to be a bit more prominent”	Increase font sizes across the interface.

Table 10: Subjects' feature requests for *SOM Browser*. Shown are response codes along with example data and interpretive summary. The number shown in parentheses behind each code denotes how many individual subjects gave responses corresponding to that code.

## 6 Discussion

In *SOM Browser*, we have developed a practical application for sound corpus organization using SOMs. To minimize empty areas in our interactive map interface, we have extended the SOM algorithm with a new method for node population, FNP, that reduces the number of empty nodes at the cost of some additional vector quantization. In a series of qualitative interviews with audio professionals we have identified an established sample library workflow, as well as confirmed participants’ interest and proclaimed need for alternate interfaces. Participants were asked to use *SOM Browser* and give feedback, with most first impressions being characterized as positive, although the organization of sounds in the map interface was seen as not easily comprehensible.

In the remaining sections, we further discuss the reported results and offer some interpretation and contextualization of them before moving on to a summary of the work presented in this thesis, including a look at its strengths, weaknesses and limitations. Finally, we suggest areas to explore in future work.

### 6.1 Interpreting the Results

Section 5.1 lays out the process of finding an optimal set of parameters for a SOM of the *Drum Essentials* sample library. Unsurprisingly, the longer the algorithm is trained, the better it performs. Noticeable is the jump in quality for both quantization errors and map emptiness when going from a training duration of  $10^4$  to  $10^5$ . These findings coincide with Kohonen (1990), whose section on “Practical Hints for the Application of the Algorithm” offers the following advice:

“Typically we have used up to 100 000 steps in our simulations, but for “fast learning”, e.g., in speech recognition, 10000 steps and even less may sometimes be enough.”

A recommendation for users of *SOM Browser* could be to find a general combination of settings that works well for their audio files with  $10^4$  steps and then run the algorithm again for  $10^5$  steps to further improve map quality. Nonetheless it would be interesting to investigate this further and see whether users actually perceive a noticeable difference when using significantly longer training times. As for the performance of our software, we see clear room for improvement of the training algorithm by porting it from JavaScript to C and making use of optimization techniques such as parallel processing and Graphics Processing Unit (GPU) acceleration.

Rather expectedly, a higher  $\alpha_{initial}$  decreases map quantization errors for our data set. This mirrors the theory behind the algorithm, according to which more drastic node adjustment in the beginning of the training phase creates the overall structure of the map, whereas at the end of training, when  $\alpha$  has decreased drastically, only minor local adjustments happen. Starting with a lower  $\alpha_{initial}$  would prohibit strong node adjustment overall. Interestingly, the choice of  $\alpha_{initial}$  does not seem to improve vector node counts or map emptiness, which appear to hover around a middle value in Figure 9.

Our finding that larger starting values for the neighborhood function radius actually produced worse results than using smaller starting radii contradicts Kohonen’s recommendations (Kohonen, 1990), who states:

“If the neighborhood is too small to start with, the map will not be ordered globally. Instead various kinds of mosaic-like parcellations of the map are seen”

Our subjective impression when exploring the map of drum sounds in *SOM Browser* does not really match this assertion. SOM global ordering could be further assessed by calculating maps for data containing group labels and examining if data from the same group is mapped to the same area of the SOM.

When comparing the three different types of learning rate factors that were implemented, the reciprocally decreasing function (“Inverse”) was outperformed by the two other methods. We believe the reason behind this to be that  $\alpha$  drops off too strongly before the map has gained its global structure. BDH on the other hand performs best out of all three methods, which most likely has to be attributed to its ability to adapt the size of  $\alpha$  locally. As to why BDH delivered the smallest median quantization error with a choice of  $m = 0$ , which effectively cancels out all adaptive local learning and results in a constant  $\alpha = \alpha_{initial}$ , we are uncertain at this time. The simulations presented in the original paper introducing the method obtain best results for other values (see Figure 2 in (Bauer et al., 1996, p.18)). It should be noted however that the authors only examine one- and two-dimensional input data.

With regard to FNP, the extension to the original SOM algorithm we implemented, we consider it a practical usability improvement at the cost of additional map distortion. Nevertheless, the enhancement of other aspects of the produced map (most importantly, a lower map emptiness) justify its application. Naturally, if more of the original map’s nodes are populated to begin with, less overall distortion will be added by applying FNP.

While the conducted interviews were analyzed qualitatively, we did gather some quantitative data through the usage of Likert scales. This was done to gauge participants first reactions to *SOM Browser* quickly and coarsely. In Figure 16, it can be seen at a glance that most answers lie on the positive side of the scales.

With the median answer to question 2.2 about participants’ satisfaction with their current options for interacting with the presented sample library being “somewhat satisfied”, and participants more in-depth responses showing some frustrations, we infer that their current workflow can be seen as good enough, but nothing more. Subjects are aware of its shortcomings and are interested in new ways of working, which attests to the need for new tools such as what was developed in this thesis. As a by-product of this interview series, a model for the current practice of working with sample libraries could be established (see Figure 17). Particularly interesting is the iterative interplay between participants mental representation of the sounds they are looking for and contextual evaluation of their search results within their production environment.

As for *SOM Browser*, interview participants’ first impressions were mostly positive, yet also critical of several aspects of the software. Clearly, potential for improvement remains. Most importantly, subjects were not able to comprehend the overall order of the map layout, which caused frustration and some fairly critical responses. We believe there to be several causes of this. It is certainly possible that the particular SOM we presented to subjects did not accurately serve as a similarity map of our sample library, perhaps due to the previously discussed fact that we used a small radius for the neighborhood function. This could have had the negative effect on the global order of the map that is

predicted by Kohonen (1990) (see the quote above). At the same time, we suspect that a methodological flaw in our question design could also have influenced participants’ inability to discern order within the map; we never suggested the map could have anything to do with any kind of potential sonic similarity. Instead we asked “What do you think about the organization of sounds in this interface?” and “What do you think the axes represent?”. Some responses to the first question did include “there seem to be areas somehow” and “Okay, so these are the percussion ...”, perhaps indicating that some local similarities could be identified, but not placed in a context concerning the entire map. But the second question is effectively impossible to answer correctly; the axes of a two-dimensional SOM are in essence meta-features, or features of features. This is not something that can be concluded without additional information. We were aware of the impossible nature of the question when designing the interview, but decided to ask it because we were interested in what possible answers we would encounter. Instead, it would have perhaps been more beneficial to position the map interface from the beginning as representing some sort of similarity and then gauging whether participants can accept it and relate to it.

Participants clearly stated their preference for their established workflow. This was to be expected, as our presented software is limited in scope and still early in its development process. Additionally, interview participants were professionals, who presumably have spent some time and effort to optimize their workflow. Nevertheless, the participants recognized some potential in *SOM Browser* and stated a desire to explore incorporating it into their workflow, under the condition that it could be integrated with their existing production environments. Participants also mentioned numerous feature requests, which could be incorporated into the software to improve usability, such as support for Arrow key navigation, ways to filter by feature and some form of additional categorization.

## 6.2 Strengths, Weaknesses and Limitations

In summarizing the present work, we can identify some strengths as well as weaknesses that bear mentioning briefly. Listed in addition are limiting factors of the work, before we move on to a discussion of potential future work and a final conclusion.

**Strengths** Our SOM implementation offers a quick, unsupervised, visual layout of a sound corpus. It can be used for completely unlabelled data and can therefore be a benefit to audio experts and more casual music producers alike. The addition of a SOM implementation to *CataRT* and *MuBu For Max* gives expert audio software users that build their own *MaxMSP* performance systems an addition to their existing toolbox. *SOM Browser* on the other hand is a simple to use, cross-platform application that allows no-frills contact with a ML algorithm that can currently not be found in commercial audio software.

**Weaknesses** It stands to reason that the collection of audio features used in this work can be improved upon, or should at least be re-evaluated critically in order to improve the quality of the produced maps. Interview participants struggled with the presented map layout, which could be due to a lack of information, but also due to insufficient quality of the SOM they saw. Mel Frequency Cepstral Coefficients (MFCCs) could be used to represent the spectral envelope of the sounds better. Additionally, each sample’s temporal envelope should be considered.

The qualitative analysis of the conducted interviews could also be improved upon by using multiple independent researchers for the coding process. As it stands, no inter-rater reliability exists, as the analysis was done by the author. This also means that although care was taken to work diligently and objectively, the potential for bias in the coding and the general interview evaluation results is obvious.

**Limitations** Although the *Drum Essentials* data set used in this work can be considered representative of a contemporary sample library, the work presented here needs to be applied to other collections of sounds as well before it can be considered more universally applicable. In conducting the interviews, we were mainly interested in getting an overview of possible responses and reactions. To this end, we chose a qualitative, semi-structured approach that prevents us from presenting conclusive, quantitative results. Finally, a limitation of *SOM Browser* is that it is not a tool for very large sample libraries; the library used by us, consisting of just over one thousand sound files, presents essentially the sensible limit of files that can be displayed.

### 6.3 Outlook

As discussed earlier, *SOM Browser* offers ample opportunity for future work and improvements. A technical improvement could be the implementation of the *Batch Map*, which offers to potentially increase speed by an order of magnitude (Kohonen and Honkela, 2007). Another welcome addition would be node initialization using PCA, as can be done in the SOM toolbox for MATLAB (Vesanto et al., 2000). Finally, a reimagining of the software in plug-in format, either as a *MaxForLive* device for Ableton Live or as a Virtual Studio Technology (VST) instrument, would be a more substantial undertaking that would satisfy the desire of our interview participants for a more direct inclusion of *SOM Browser* functionality in their existing production environments.

For the SOM algorithm as a whole and its utility for sound corpus organization, an exploration of input dimension weightings for different applications could be interesting. Perhaps the same set of features can be used with different weightings depending on if the corpus contains percussive, monophonic or polyphonic sounds. Additionally, an augmentation of the audio feature input data with file metadata could also be considered. One can imagine for example using a string distance function to compute file name similarity.

### 6.4 Conclusion

While the availability of large sample libraries and cheap storage solutions has increased, tools for navigating, searching and organizing these increasingly unmanageable audio file collections have not kept pace. The present thesis approaches this problem from a practical perspective. We implement the Self-Organizing Map algorithm and apply it to sound corpus organization. Our implementation is embedded in two applications. In *MaxMSP*, it can be used to extend *CataRT* and the *MuBu For Max* package, offering new possibilities to expert users. In *SOM Browser*, it is used for a fast, visual interface for sample library exploration, offering an alternative to the established music production workflow. In doing this, our software offers direct access to a ML algorithm that to our knowledge is not available in any commercial audio software. We also extend the core algorithm with FNP,

a method that allows an improved utilization of the screen space allocated for the SOM interface by reducing the areas of the map to which no sound is mapped. Although the presented work focuses on utilizing short recordings of drum sounds, it could in principle be applied to other sounds as well.

After searching for optimal parameters according to objective measures of map quality and producing a map for a chosen corpus of drum sounds, we conduct a series of qualitative interviews with audio professionals. Participants' responses allow us to identify a prevalent method of working with sample libraries, which we codify into a generalized model of the established workflow. We confirm the need for and interest in alternate interfaces and present participants with our developed software. Interview feedback was sufficiently positive to warrant further work towards this end.



## 7 References

- Ableton AG (2019a): *Ableton Live 10*. Software. URL <https://www.ableton.com/en/live/>. Access 7.2.2019.
- Ableton AG (2019b): *Drum Essentials*. Online. URL <https://www.ableton.com/en/packs/drum-essentials/>. Access 7.2.2019.
- Adeney, Roland and Andrew R Brown (2009): “Performing with grid music systems.” In: *Improvise: The Australasian Computer Music Conference 2009*. Australasian Computer Music Association (ACMA), pp. 102–110.
- Algonaut (2019): *Atlas*. Software. URL <https://www.algonaut.tech/>. Access 7.2.2019.
- Barbiero, Phillip (2017): *pbarbiero/basic-electron-react-boilerplate: Modern and Minimal Electron + React Starter Kit*. Online. URL <https://github.com/pbarbiero/basic-electron-react-boilerplate>. Access 7.2.2019.
- Bauer, H.-U.; Ralf Der; and Michael Herrmann (1996): “Controlling the magnification factor of self-organizing feature maps.” In: *Neural computation*, **8**(4), pp. 757–771.
- Coleman, Graham (2007): “Mused: Navigating the Personal Sample Library.” In: *ICMC*. Citeseer.
- Cosi, Piero; Giovanni De Poli; and Giampaolo Lauzzana (1994): “Auditory Modelling and Self-Organizing Neural Networks for Timbre Classification.” In: *Journal of New Music Research*, **23**(1), pp. 71–98.
- Cycling '74 (2019): *Max*. Software. URL <https://cycling74.com/>. Access 7.2.2019.
- de la Cuadra, Patricio (2019): “Pitch Detection Methods Review.” URL <https://ccrma.stanford.edu/~pdelac/154/m154paper.htm>.
- Facebook (2019): *React*. Software. URL <https://reactjs.org/>. Access 7.2.2019.
- Fletcher, Harvey and Wilden A Munson (1933): “Loudness, its definition, measurement and calculation.” In: *Bell System Technical Journal*, **12**(4), pp. 377–430.
- Font, Frederic; Gerard Roma; and Xavier Serra (2013): “Freesound technical demo.” In: *Proceedings of the 21st ACM international conference on Multimedia*. ACM, pp. 411–412.
- Fried, Ohad; Zeyu Jin; Reid Oda; and Adam Finkelstein (2014): “AudioQuilt: 2D Arrangements of Audio Samples using Metric Learning and Kernelized Sorting.” In: *NIME*. pp. 281–286.
- Garland, Ron (1991): “The mid-point on a rating scale: Is it desirable.” In: *Marketing bulletin*, **2**(1), pp. 66–70.
- Gillet, Olivier and Gaël Richard (2006): “ENST-Drums: an extensive audio-visual database for drum signals processing.” In: *ISMIR*. pp. 156–159.
- GitHub (2019): *Electron*. Software. URL <https://electronjs.org/>. Access 7.2.2019.

- Goto, Masataka; Hiroki Hashiguchi; Takuichi Nishimura; and Ryuichi Oka (2002): “RWC Music Database: Popular, Classical and Jazz Music Databases.” In: *ISMIR*, vol. 2. pp. 287–288.
- Heise, Sebastian; Michael Hlatky; and Jörn Loviscach (2008): “Soundtorch: Quick Browsing in Large Audio Collections.” In: *Audio Engineering Society Convention 125*. Audio Engineering Society.
- Henwood, Karen and Nick Pidgeon (2003): *Grounded theory in psychological research*. American Psychological Association, pp. 131–55.
- Iced Audio (2019): *Audio Finder*. Software. URL <http://www.icedaudio.com/>. Access 7.2.2019.
- Knees, Peter; et al. (2015): “Two Data Sets for Tempo Estimation and Key Detection in Electronic Dance Music Annotated from User Corrections.” In: *ISMIR*. pp. 364–370.
- Kohonen, Teuvo (1990): “The Self-Organizing Map.” In: *Proceedings of the IEEE*, **78**(9), pp. 1464–1480.
- Kohonen, Teuvo (2001): *Self-Organizing Maps*, vol. 30 of *Springer Series in Information Sciences*. Heidelberg: Springer.
- Kohonen, Teuvo (2005): “The Self-Organizing Map (SOM).” URL <http://www.cis.hut.fi/somtoolbox/theory/somalgorithm.shtml>.
- Kohonen, Teuvo and Timo Honkela (2007): “Kohonen network.” URL [http://www.scholarpedia.org/article/Kohonen\\_network](http://www.scholarpedia.org/article/Kohonen_network).
- Lazar, Jonathan; Jinjuan Heidi Feng; and Harry Hochheiser (2017): *Research methods in human-computer interaction*. Morgan Kaufmann.
- Lerch, Alexander (2012): *An introduction to audio content analysis: Applications in signal processing and music informatics*. Wiley-IEEE Press.
- Lykartsis, Athanasios (2014): *Evaluation of accent-based rhythmic descriptors for genre classification of musical signals*. Master’s thesis, Master’s thesis, Audio Communication Group, Technische Universität Berlin . . . .
- Maaten, Laurens van der and Geoffrey Hinton (2008): “Visualizing data using t-SNE.” In: *Journal of machine learning research*, **9**(Nov), pp. 2579–2605.
- Mathieu, Benoit; Slim Essid; Thomas Fillon; Jacques Prado; and Gaël Richard (2010): “YAAFE, an Easy to Use and Efficient Audio Feature Extraction Software.” In: *ISMIR*. pp. 441–446.
- McDonald, Kyle and Manny Tan (2019): *The Infinite Drum Machine*. Online. URL <https://experiments.withgoogle.com/drum-machine>. Access 7.2.2019.
- Merenyi, Erzsébet; Abha Jain; and Thomas Villmann (2007): “Explicit magnification control of self-organizing maps for “forbidden” data.” In: *IEEE Transactions on Neural Networks*, **18**(3), pp. 786–797.

- Moffat, David; David Ronan; Joshua D Reiss; et al. (2015): “An evaluation of audio feature extraction toolboxes.” In: *18th International Conference on Digital Audio Effects (DAFx)*.
- Moore, Brian CJ; Brian R Glasberg; and Thomas Baer (1997): “A model for the prediction of thresholds, loudness, and partial loudness.” In: *Journal of the Audio Engineering Society*, **45**(4), pp. 224–240.
- Nielsen, Bjørn Næsby (2019): *Sononym*. Software. URL <https://www.sononym.net/>. Access 7.2.2019.
- Oppenheim, Alan V and Ronald W Schaffer (2014): *Discrete-time signal processing*. Pearson Education.
- Pampalk, Elias; Peter Hlavac; and Perfecto Herrera (2004): “Hierarchical organization and visualization of drum sample libraries.” In: *7th International Conference on Digital Audio Effects (DAFx)*. pp. 378–383.
- Peeters, Geoffroy (2004): *A large set of audio features for sound description (similarity and classification) in the CUIDADO project*. Tech. rep., IRCAM.
- Rawlinson, Hugh; Nevo Segal; and Jakub Fiala (2015): “Meyda: an audio feature extraction library for the web audio api.” In: *The 1st Web Audio Conference (WAC)*. Paris, Fr.
- Rawlinson, Hugh; Nevo Segal; and Jakub Fiala (2019a): “Meyda: Audio feature extraction for JavaScript.” URL <https://meyda.js.org/audio-features>.
- Rawlinson, Hugh; Nevo Segal; and Jakub Fiala (2019b): “Meyda: Audio feature extraction for JavaScript.” URL <https://github.com/meyda/meyda>.
- Saldaña, Johnny (2015): *The coding manual for qualitative researchers*. Sage.
- Schnell, Norbert; et al. (2009): “MuBu and friends—assembling tools for content based real-time interactive audio processing in Max/MSP.” In: *ICMC*.
- Schnell, Norbert; et al. (2019a): *MuBu for Max - A toolbox for Multimodal Analysis of Sound and Motion, Interactive Sound Synthesis and Machine Learning*. Online. URL <http://ismm.ircam.fr/mubu/>. Access 7.2.2019.
- Schnell, Norbert; et al. (2019b): *MuBu for Max - A toolbox for Multimodal Analysis of Sound and Motion, Interactive Sound Synthesis and Machine Learning*. Online. URL <http://forumnet.ircam.fr/product/mubu-en/>. Access 7.2.2019.
- Schwartz, Barry (2004): *The paradox of choice: Why more is less*, vol. 6. HarperCollins New York.
- Schwarz, Diemo; Grégory Beller; Bruno Verbrugghe; and Sam Britton (2006): “Real-time corpus-based concatenative synthesis with catart.” In: *9th International Conference on Digital Audio Effects (DAFx)*. pp. 279–282.

- Shier, Jordie; Kirk McNally; and George Tzanetakis (2017): “Analysis of Drum Machine Kick and Snare Sounds.” In: *Audio Engineering Society Convention 143*. Audio Engineering Society.
- Torvalds, Linus and Junio Hamano (2019): *Git*. Software. URL <https://git-scm.com/>. Access 7.2.2019.
- TU Berlin, Ethik-Kommission (2019): *Ethik-Kommission*. URL [https://www.ipa.tu-berlin.de/menue/einrichtungen/gremienkommissionen/ethik\\_kommission/](https://www.ipa.tu-berlin.de/menue/einrichtungen/gremienkommissionen/ethik_kommission/).
- Vagias, Wade M (2006): “Likert-type Scale Response Anchors. Clemson International Institute for Tourism.” In: *8 Research Development, Department of Parks, Recreation and Tourism Management, Clemson University*.
- Vesanto, Juha; Johan Himberg; Esa Alhoniemi; and Juha Parhankangas (2000): “SOM toolbox for Matlab 5.” In: *Helsinki University of Technology, Finland*, p. 109.
- World Wide Web Consortium (W3C) (2019): *Web Audio API*. Online. URL <https://www.w3.org/TR/webaudio/>. Access 7.2.2019.
- Yin, Hujun (2007): “Nonlinear dimensionality reduction and data visualization: a review.” In: *International Journal of Automation and Computing*, **4**(3), pp. 294–303.
- Zwicker, Eberhard (1961): “Subdivision of the audible frequency range into critical bands (Frequenzgruppen).” In: *The Journal of the Acoustical Society of America*, **33**(2), pp. 248–248.

# Appendices

All supplementary materials listed here are included with the digital submission of this work.

## A Source Code

The L<sup>A</sup>T<sub>E</sub>X sources for this work are located in the directory `tex/`.

The source code for the applications that were developed for this thesis can be found in the following directories:

*mubu-SOM-js*: `dev/mubu-som-js/`

*SOM Browser*: `dev/som-browser/`

## B Sample Library

The *Drum Essentials* sample library that was used as a sound corpus can be found in `evaluation/DrumEssentials/`.

## C Evaluation Map Data

All maps used for the results presented in Section 5 can be found in `evaluation/eval_maps/`.

## D Interview Data

The recorded interviews, their transcriptions and a spreadsheet with the codified data are located in `evaluation/interviews/`.

## E MATLAB Figures

The MATLAB scripts to create the figures shown in Section 5 can be found in `evaluation/matlab/`.

## Acronyms

**ACA** Audio Content Analysis.

**API** Application Programming Interface.

**BDH** Bauer-Der-Herrmann Algorithm.

**BMU** Best Matching Unit.

**DAW** Digital Audio Workstation.

**FFT** Fast Fourier Transform.

**FNP** Forced Node Population.

**GPU** Graphics Processing Unit.

**GUI** Graphical User Interface.

**HCI** Human-Computer Interaction.

**IRCAM** Institut de recherche et coordination acoustique/musique.

**MFCC** Mel Frequency Cepstral Coefficient.

**MIR** Music Information Retrieval.

**ML** Machine Learning.

**NaN** Not a Number.

**OS** Operating System.

**PCA** Principal Component Analysis.

**RAM** Random-Access Memory.

**SOM** Self-Organizing Map.

**TU Berlin** Technische Universität Berlin.

**VST** Virtual Studio Technology.

## List of Figures

1	mubu-SOM-js . . . . .	10
2	<i>CataRT</i> : with and without SOM . . . . .	11
3	<i>CataRT</i> : SOM with added noise . . . . .	11
4	<i>SOM Browser</i> . . . . .	15
5	<i>SOM Browser</i> without audio files loaded . . . . .	16
6	<i>SOM Browser</i> : Mock-up outlining system states . . . . .	18
7	Comparison of maps with (7b) and without (7a) FNP . . . . .	23
8	Effects of Training Duration on SOM Metrics . . . . .	32
9	Effects of $\alpha_{initial}$ on SOM Metrics . . . . .	32
10	Effects of $r_{start}$ on SOM Metrics . . . . .	33
11	Effects of $r_{end}$ on SOM Metrics . . . . .	34
12	Comparison of SOM learning rate types . . . . .	34
13	BDH: Effect of Magnification Control $m$ . . . . .	35
14	Comparison of FNP influence on SOM . . . . .	36
15	<i>SOM Browser</i> : Visual Influence of FNP . . . . .	36
16	Interview Ratings . . . . .	37
17	Established sample library workflow . . . . .	40

## List of Listings

1	dev/mubu-som-js/descriptor_som.js: createSOM() . . . . .	13
2	dev/mubu-som-js/descriptor_som.js: training() . . . . .	13
3	dev/mubu-som-js/descriptor_som.js: Neuron position updates . . . . .	13
4	dev/mubu-som-js/descriptor_som.js: BMU identification . . . . .	14
5	dev/som-browser/src/components/App.js: handleAnalyzeClick() [ex- cerpt] . . . . .	20
6	dev/som-browser/src/background/extractFeatures.js: Audio feature ex- traction . . . . .	20
7	dev/som-browser/src/components/App.js: GUI Components . . . . .	21
8	FNP implementation . . . . .	23



## List of Tables

1	mubu-SOM-js: Messages for algorithm control . . . . .	12
2	<i>Drum Essentials</i> : Instrument category counts . . . . .	25
3	Comparison of SOM Learning Rate Types . . . . .	35
4	Influence of FNP . . . . .	36
5	Established Sample Library Workflow Description: Response Codes . . . . .	38
6	Established Sample Library Workflow Assessment: Response Codes . . . . .	39
7	<i>SOM Browser</i> : Positive Responses . . . . .	41
8	<i>SOM Browser</i> : Negative Responses . . . . .	41
9	Workflow Comparison: <i>SOM Browser</i> vs. Established Workflow . . . . .	42
10	<i>SOM Browser</i> : Feature Requests . . . . .	43