Technische Universität Berlin

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

Self-Organizing Maps for Sound Corpus Organization

MASTER'S THESIS

Vorgelegt von: Jonas Margraf
Matrikelnummer: 372625
E-Mail: jonasmargraf@me.com

Erstgutachter: Prof. Dr. Stefan Weinzierl
Zweitgutachter: Dr. Diemo Schwarz
Datum: February 10, 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 February 10, 2019

....
Jonas Margraf

Abstract An english abstract.

Zusammenfassung Die Zusammenfassung auch auf Deutsch.

Acknowledgements

This is where the thank yous go.

Contents

1 Introduction

This is the Introduction. Here's a citation about Self-Organizing Maps (SOMs)(Kohonen, 1990).

1.1 Motivation and Problem Description

1.2 Aims and Objectives

1.3 Previous Work

2 Background

This is the Background section.

2.1 Audio Feature Extraction

Make sure to quote Lerch (2012), Rawlinson et al. (2015), Rawlinson et al. (2019a), Mathieu et al. (2010) Mathieu et al. (2019).

2.1.1 Fundamentals

2.1.2 Audio Pre-Processing

2.1.3 Time-Domain Features

Define $x[n]$, n

2.1.3.1 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., 2019b) 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

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

2.1.3.2 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

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

2.1.4 Frequency-Domain Features

Define N_{FFT}, $X(k)$

2.1.4.1 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

$$v_{SC} = \frac{\sum\limits_{k=0}^{N_{FFT}/2-1} k \cdot |X(k)|^2}{\sum\limits_{k=0}^{N_{FFT}/2-1} |X(k)|^2}. \tag{3}$$

2.1.4.2 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

$$v_{SFL} = \frac{\sqrt[N_{FFT}/2]{\prod\limits_{k=0}^{N_{FFT}/2-1} |X(k)|}}{(2/N_{FFT}) \cdot \sum\limits_{k=0}^{N_{FFT}/2-1} |X(k)|}. \tag{4}$$

2.1.4.3 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

$$v_{SKU} = \frac{2 \sum\limits_{k=0}^{N_{FFT}/2-1} (|X(k)| - \mu_{|X|})^4}{N_{FFT} \cdot \sigma_{|X|}^4} - 3, \tag{5}$$

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

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

$$v_{SSK} = \frac{2 \sum\limits_{k=0}^{N_{FFT}/2-1} (|X(k)| - \mu_{|X|})^3}{N_{FFT} \cdot \sigma_{|X|}^3}. \tag{6}$$

2.1.4.5 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

$$v_{SSL} = \frac{\sum_{k=0}^{N_{FFT}/2-1} (k - \mu_k)(|X(k)| - \mu_{|X|})}{\sum_{k=0}^{N_{FFT}/2-1} (k - \mu_k)^2}. \tag{7}$$

2.1.4.6 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

$$v_{SSP} = \frac{\sum_{k=0}^{N_{FFT}/2-1} (k - v_{SC})^2 \cdot |X(k)|^2}{\sum_{k=0}^{N_{FFT}/2-1} |X(k)|^2}. \tag{8}$$

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

$$v_{SR} = i \Bigg|_{\sum_{k=0}^{i} |X(k)| = \kappa \cdot \sum_{k=0}^{N_{FFT}/2-1} |X(k)|}. \tag{9}$$

2.1.5 Perceptual Features

2.1.5.1 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

$$v_{TL} = \sum_{i=1}^{24} v_{SL}(i), \tag{10}$$

where

$$v_{SL}(i) = E(i)^{0.23} \tag{11}$$

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

2.2 Self-Organzing 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 d-dimensional weight vector $m = [m_1, ..., m_d]$, where d 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).
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).

what is it used for? pattern recognition

mathematical definition

algorithm described using pseudo-code
initialization
input data normalization

3 Implementation

This is the Implementation.

3.1 Groundwork: CataRT Extension

3.2 SOM Browser

4 Evaluation

This is the Evaluation.

4.1 Measuring SOM-Induced Quantization

4.2 Online Sound Similarity Survey

4.3 Semistructured User Interviews

5 Results

This is the Results section.

6 Discussion

This is the Discussion.

6.1 Outlook

7 References

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.

de la Cuadra, Patricio (2019): "Pitch Detection Methods Review." URL `https://ccrma.stanford.edu/~pdelac/154/m154paper.htm`.

DeSieno, Duane (1988): "Adding a conscience to competitive learning." In: *IEEE international conference on neural networks*, vol. 1. Institute of Electrical and Electronics Engineers New York, pp. 117–124.

Fasciani, Stefano (2016): "TSAM: a tool for analyzing, modeling, and mapping the timbre of sound synthesizers." In: .

Fiebrink, Rebecca and Baptiste Caramiaux (2016): "The machine learning algorithm as creative musical tool." In: *Handbook of Algorithmic Music*.

Fletcher, Harvey and Wilden A Munson (1933): "Loudness, its definition, measurement and calculation." In: *Bell System Technical Journal*, **12**(4), pp. 377–430.

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.

Gillet, Olivier and Gaël Richard (2006): "ENST-Drums: an extensive audio-visual database for drum signals processing." In: *ISMIR*. pp. 156–159.

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.

Kohonen, Teuvo (1990): "The Self-Organizing Map." In: *Proceedings of the IEEE*, **78**(9), pp. 1464–1480.

Kohonen, Teuvo (1997): "Exploration of very large databases by self-organizing maps." In: *Proceedings of International Conference on Neural Networks (ICNN'97)*, vol. 1. IEEE, pp. PL1–PL6.

Kohonen, Teuvo (2001): *Self-Organizing Maps*, vol. 30 of *Springer Series in Information Sciences*. Heidelberg: Springer.

Kohonen, Teuvo and Timo Honkela (2007): "Kohonen network." URL `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 ....

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.

Mathieu, Benoit; Slim Essid; Thomas Fillon; Jacques Prado; and Gaël Richard (2019): "Yaafe - audio features extraction." URL `http://yaafe.sourceforge.net/`.

Mayring, Philipp (2010): "Qualitative Inhaltsanalyse." In: *Handbuch qualitative Forschung in der Psychologie*. Springer, pp. 601–613.

Merenyi, Erzsbet; 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: .

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.

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://github.com/meyda/meyda`.

Rawlinson, Hugh; Nevo Segal; and Jakub Fiala (2019b): "Meyda: Audio feature extraction for JavaScript." URL `https://meyda.js.org/audio-features`.

Scholler, Simon and Hendrik Purwins (2010): "Sparse coding for drum sound classification and its use as a similarity measure." In: *Proceedings of 3rd international workshop on Machine learning and music*. ACM, pp. 9–12.

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.

Vagias, Wade M (2006): "Likert-type Scale Response Anchors. Clemson International Institute for Tourism." In: *& 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.

Villmann, Thomas and Jens Christian Claussen (2006): "Magnification control in self-organizing maps and neural gas." In: *Neural Computation*, **18**(2), pp. 446–469.

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

A LaTeX Sources

The LaTeX sources for this work can be found in XXX.

B Thesis Bibliography

The references used in this work can be found in XXX.

Acronyms

SOM Self-Organizing Map.

List of Figures

List of Listings

List of Tables

Digital Resource

This page holds a data disk.