Music Information Retrieval and Recommendation

Christian Tobias (ct042) | Matrikel-Nr. 42791
Hochschule der Medien Stuttgart | Wintersemester 2021/22
Aktuelle Themen (AM3) | Prof. Dr. Andreas Koch

Abstract

Music Information Retrieval bildet die Basis vieler automatisierter und KI-gestützter Systeme im Audiobereich. Durch den Anstieg der Popularität von Streaming Services wie Spotify, Apple Music, etc. hat der Bedarf für präzise Musik-Klassifizierung und personalisierte Empfehlung deutlich zugenommen. Dieser Teilbereich der Music Information Retrieval ist besonders eng verknüpft mit Künstlicher Intelligenz. Dieses White-Paper beleuchtet die Funktionsweisen der grundlegenden Technologien und Prinzipien der Music Information Retrieval, die für Musik-Klassifizierung notwendig sind. Weitergehend werden darauf aufbauende Musikempfehlungsalgorithmen genauer betrachtet.

1 Einführung

Music Information Retrieval (MIR) ist eine interdisziplinäre Forschungsrichtung, die sich mit den Methoden, Prozessen und Systemen zur Extraktion von Informationen aus Musik befasst. [1] Die Musik kann dabei in symbolischen Formaten (z.B. MIDI-Dateien), in einem Vektor-Format (z.B. Noten) oder in einer digitalen Audio-Format (z.B. MP3-Dateien) repräsentiert sein. MIR nutzt die Technologien und Methoden aus der Audiosignalverarbeitung, maschinellem Lernen, Database-Management,

human-computer Interaction oder Musiksoziologie. [2]

Die typischen Aufgabenfelder der Music Information Retrieval sind beispielsweise die Erkennung und Separation von beteiligten Instrumenten, Elementen oder Bestandteilen in der Musik, die Detektion von musikalischen Schlägen, Tempi und Taktarten oder Tonhöhen, Melodien und Tonarten.

Weiterführend kann die Struktur, das Genre und der „*Mood*“ eines Songs ermittelt werden. Eine weitere Aufgabe ist der Transfer zwischen verschiedenen Repräsentationsmodi der Musik, beispielsweise der Transkription, also dem Übertragen von Musik im Audioformat zu einem Vektorformat. [3] Ein weiterer Bestandteil des MIR ist das sogenannte Fingerprinting, bei dem jedem Stück Musik ein eindeutiger digitaler „Fingerabdruck“ zugewiesen wird. Genauer, mittels MIR können also Algorithmen erschaffen werden, die beispielsweise das Original zu einem Cover-Song zuordnen, Musikstücke mit Noten synchronisieren oder Lieder erkennen können, die lediglich vorgesummt oder -gesungen werden. [4] In dem Bereich des MIR, insbesondere im Teilbereich der Musikempfehlungsalgorithmen, hat die Relevanz von KI-basierten Systemen durch den technologischen Fortschritt in den letzten Jahren deutlich zugenommen. [5]

2 MIR-Systeme

Ein übliches MIR-System nimmt am Eingang ein Stück Musik entgegen, meistens in Form einer Audiodatei, welches in der Folge die folgenden drei Schritte durchläuft:

- Segmentation
- Feature Extraction
- Machine Learning

Am Ausgang eines solchen Systems steht dann die erzielte musikalische Information. [4] Auf die einzelnen Schritte wird in der Folge noch genauer eingegangen.

2.1 Segmentation

Der Schritt der Segmentation beschäftigt sich damit, die Musik in sinnvolle Abschnitte unterteilen. Dabei gibt es mehrere Ansätze. Je nach Zielsetzung kann dabei zwischen musikalisch und technisch orientierten Segmentierungsmethoden ausgewählt werden. Eine musikalisch orientierte Methode ist zum Beispiel die Beat-Segmentation, bei der die Musik mit u.a. einer

Transienten-Erkennung in musikalische Schläge unterteilt wird. [6]

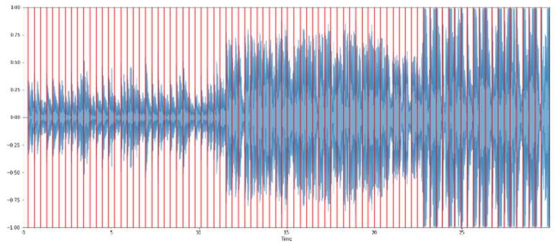

Abbildung 1 *Beispiel einer Beat-Segmentation*

Alternativ kann auch eine sogenannte Onset-Segmentation zum Einsatz kommen, die Noten-Onsets erkannt und die Audiodatei dementsprechenden unterteilt, sodass am Anfang jedes Segments ein Notenevent stattfindet.

Eine technisch orientierte Segmentierungsmethode ist das Fenstern, bei dem die Audiodatei in gleich große Abschnitte unterteilt wird, die die Signalverarbeitung erleichtern (beispielsweise Abschnitte von 2048 Audio-Samples).

2.2 Feature Extraction

Die Feature Extraction hat das Ziel, Muster in der Musik sichtbar zu machen, die für die menschliche Wahrnehmung relevant sind. Dabei unterscheidet man zwischen Low, Mid und High-Level Features.

Low-Level Features können als Kurzeitdeskriptoren aus dem Audiosignal bestimmt werden. Das umfasst Parameter wie Klangfarbe, zeitliche Verläufe, Intensität, Tonhöhenverläufe. Mithilfe der Low-Level Features und Algorithmen ist es in der Folge möglich musikalische Eigenschaften wie Tempo, Taktart, Harmonie- und Melodieverläufe als Mid-Level Features aus dem Audiosignal zu extrahieren. [7] Diese extrahierten Deskriptoren können für das Training von neuronalen Netzen im nächsten Schritt eingesetzt werden, um Parameter mit einer höheren semantischen Bedeutung (High-Level Features) wie beispielsweise ausgedrückte Emotionen vorherzusagen. Aus Low-Level Features können also Mid-Level Features ermittelt werden, welche zur Vorhersage von High-Level Features mittels KI-Systemen dienen. [8]

Realisiert werden diese Extrahierungsmethoden mit den Werkzeugen der Audiosignalverarbeitung. Ein weit verbreitetes und nützliches Feature ist das Mel-Spectogram.

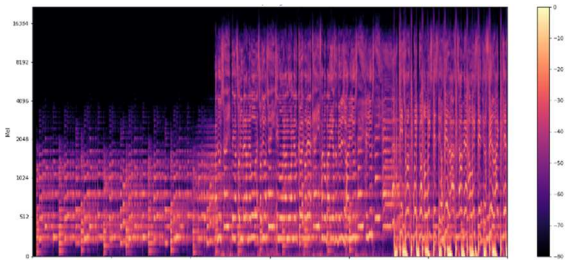

Abbildung 2 *Beispiel für ein Mel-Spectogram*

Dafür wird der spektrale Verlauf der Musik gegen die Zeit geplottet, wobei die Frequenzen in die Mel-Skala transferiert werden, welche der menschlichen Tonhöhen-Wahrnehmung besser repräsentiert. [9] Für eine effektivere Weiterverwendung können außerdem gewissen Elemente in der Musik voneinander separiert werden. Eine übliche Unterteilung ist die *Harmonic-Percussive Source Separation*. Dabei werden die harmonisch, tonalen Elemente von den perkussiven, Transienten-reichen Bestandteilen getrennt. Das kann dabei helfen, die weiteren Bearbeitungsschritte effizienter zu gestalten. So kann man aus dem harmonischen Teil beispielsweise besser die Tonart ermitteln und aus dem perkussiven Teil etwa das Tempo.

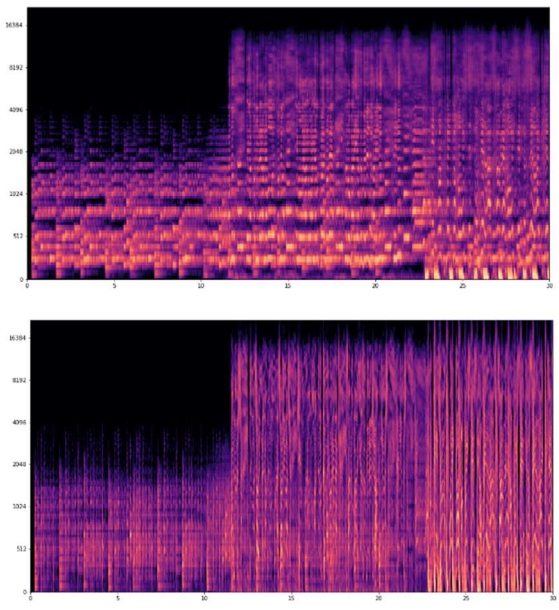

Abbildung 3 *harmonischer Bestandteil (oben) und perkussiver Bestandteil (unten)*

Ein weitere elementare Repräsentationsart, die aus der Audiosignalverarbeitung kommt, ist das Cepstrum. Das Cepstrum ist das Ergebnis einer mathematischen Transformation des Spektrums. Es wird verwendet, um periodische Strukturen im Frequenzspektrum eines Signals zu erkennen (z.B. Obertöne oder Reflexionen).[10] Bevorzugt wird dafür das Mel-Frequenz-Spektrum verwendet, um die sogenannten *Mel-Frequency Cepstral Coefficients* (kurz MFCC) zu erhalten. Diese bilden sehr kompakt viele für die menschliche Wahrnehmung relevante Information ab und sind daher elementar für MIR-Systeme. MFCCs bilden oftmals die Datengrundlage für maschinelles Lernen.

2.3 Machine Learning

Auch im Bereich des MIR kommt KI regelmäßig zum Einsatz. Insbesondere neuronale Netze mit einer oder mehreren Faltungsschichten, sogenannte Faltungsnetze oder auch Convolutional Neural Networks (CNNs), sind weit verbreitet. [11] CNNs sind besonders effektiv bei der Mustererkennung in Bilddateien und da durch den Schritt der Feature Extraction die Audiodateien in ein Bildformat übertragen wurden, bieten sich CNNs auch für

den Einsatz in einem MIR-System an.

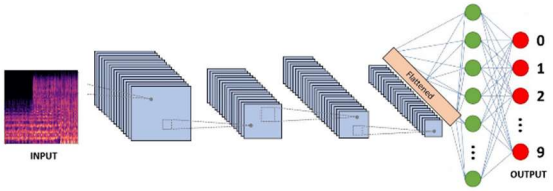

Abbildung 4 *Beispiel eines CNN*

Ein Faltungsnetz besitzt neben der Eingabe-, Ausgabe- und den versteckten Schicht noch eine oder mehrere Faltungsschichten, die wie ein Filter funktionieren. Die Eingabe wird dafür mit einer Filtermatrix gefaltet, welche spezielle Eigenschaften im Signal hervorheben soll. [12] Diese Filtermatrizen sind oftmals ebenfalls maschinell erlernt. Bei einem MIR-System dienen diese Filtermatrizen dazu, bestimmte Eigenschaften in der Musik zu erkennen, wie beispielsweise Vibrato im Gesang, Terzharmonien oder Bass Drum Hits. [13] Diese High-Level-Informationen können wiederum beispielsweise zur Einschätzung des Genres der Musik verwendet werden.

3 Recommendation

Moderne Musikempfehlungsalgorithmen machen sich drei verschiedene Prozesse zu Nutze. Jeder dieser Prozesse arbeitet wiederum mit einer Form von KI: [13]

- *Convolutional Neural Networks*
- *Collaborative Filtering*
- *Natural Language Processing*

Die Grundlage bilden dabei MIR-Systeme zur Klassifizierung der Musik. Wie im vorherigen Kapitel beschrieben benutzte diese Faltungsnetze, welche mittels Low- und Mid-Level Features der Audiodatei das Genre oder ausgedrückte Emotionen in einem Musikstück ermitteln. Im Falle von Spotify wird beispielsweise die Tonhöhenverlauf, das Timbre, die Loudness, das Tempo, die Tonart etc. bestimmt, um daraus mittels Faltungsnetzen eine Vorhersage für die Rezeption des Musikstückes treffen zu können. [14]

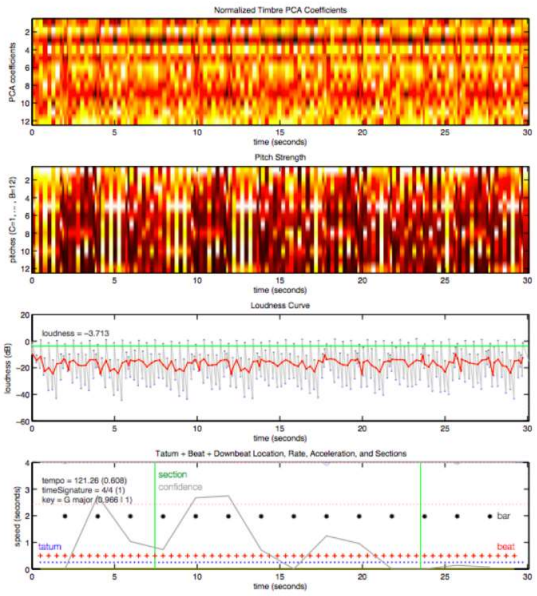

Abbildung 5 *Analyse des Songs "Around the World" von Spotify [15]*

Zusätzlich zu der Informationsgewinnung aus der Musik selbst, wenden moderne Musikempfehlungsalgorithmen zusätzlich noch zwei weitere Methoden an, um besser automatisiert empfehlen zu können.

Eine dieser Methoden ist das *Collaborative Filtering.* Die Idee dahinter ist es, die Präferenzen der User*innen aus vorhergegangenem Nutzerverhalten zu determinieren. [13] Zum Beispiel, wenn zwei Nutzer*innen zum Großteil die gleichen Musikstücke anhören, sind ihre Präferenzen wahrscheinlich ähnlich. Außerdem, wenn zwei Musikstücke von der gleichen Gruppe an Nutzer*innen gehört werden, dann klingen sie wahrscheinlich ähnlich. Das größte Problem dabei ist, dass erstmal ein Nutzerverhalten vorliegen muss, um entsprechende Daten erheben zu können und Empfehlungen aussprechen zu können. Neue oder unbekannte Musikstücke werden dabei also schlechter erfasst.

Eine weitere Methode ist das *Natural Language Processing.* Dabei werden Veröffentlichungen über Musikstücke analysiert und Stücke, die mit einem ähnlichen Vokabular beschrieben werden, miteinander gruppiert. [15]

Die Kombination aus den Daten dieser drei Methoden bildet das Fundament für moderne Empfehlungsalgorithmen, die zuverlässig und flexibel sein müssen.

4 Schlussbetrachtung

Die Popularität von KI-basierenden Systemen und Prozessen ist auch in der Welt der Musik angekommen. Dieses Paper hat eine der grundlegenden Weisen, wie KI und MIR miteinander verknüpft sind, aufgezeigt. Es wurden die wesentlichen Schritte eines MIR-Systems dargelegt. Jedes Musikstück muss dabei zunächst sinnvoll segmentiert werden, damit anschließend Low- und Mid-Level Informationen aus der Musikdatei extrahiert werden können. Diese Informationen können im Anschluss verwendet werden, um mittels KI, meistens in der Form von Faltungsnetzen, High-Level Informationen zu gewinnen.

Dieser Prozess ist ein wichtiger Bausteinen in Musikempfehlungsalgorithmen. Diese werden durch *Collaborative Filtering* und *Natural Language Processing* ergänzt, welche sich mit dem Nutzerverhalten und dem sprachlichen Beschreibungen von Musikstücken

beschäftigen, wodurch sich die Performance deutlich verbessert.

5 Literatur

1. Müller M (2015) Fundamentals of music processing. Audio, analysis, algorithms, applications. Springer, Cham, Heidelberg, New York, Dordrecht, London
2. Laurier C, Herrera P (2009) Automatic Detection of Emotion in Music. In: Vallverdú J, Casacuberta D (Hrsg) Handbook of Research on Synthetic Emotions and Sociable Robotics. IGI Global, S 9–33
3. Stigge, Roland (2003) Automatische Musiktranskription (AMT)
4. Tjoa, Steve (2014) Music Information Retrieval using Scikit-learn. https://www.youtube.com/watch?v=oGGVvTgHMHw&ab_channel=DataCouncil. Zugegriffen: 14. März 2022
5. Li T, Ogihara M (2006) Toward intelligent music information retrieval. IEEE Trans. Multimedia 8(3):564–574. doi:10.1109/TMM.2006.870730
6. Li W, Zhang X, Wang Z (2013) Music content authentication based on beat segmentation and fuzzy classification. J AUDIO SPEECH MUSIC PROC. 2013(1). doi:10.1186/1687-4722-2013-11
7. Tosta J (2021) Design und Implementierung künstlicher neuronaler Netze zur Vorhersage des semantisch-emotionalen Ausdrucks von Musik
8. Downie JS (2003) Music information retrieval. Annual Review of Information Science and Technology (37):295–340
9. Pfister B, Kaufmann T (2008) Sprachverarbeitung. Grundlagen und Methoden der Sprachsynthese und Spracherkennung. Springer-Lehrbuch. Springer, Berlin, Heidelberg
10. Wikipedia (2021) Cepstrum. https://de.wikipedia.org/wiki/Cepstrum#:~:text=Das%20Cepstrum%20(Plural%20Cepstra)%20ist,Bogert%2C%20Healy%20und%20Tukey%20eingef%C3%BChrt. Zugegriffen: 16. März 2022
11. Viswanathan AP (2016) Music Genre Classification. IJECS. doi:10.18535/ijecs/v4i10.38
12. Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P (2011) Natural Language Processing (almost) from Scratch

13. Dieleman S (2014) Recommending Music on Spotify with Deep Learning. https://benanne.github.io/2014/08/05/spotify-cnns.html. Zugegriffen: 09. März 2022
14. Roisin Loughran, Jacqueline Walker, Michael O'Neill, Marion O'Farrell The Use of Mel-frequency Cepstral Components in Musical Instrument Identification
15. Boyd C (2019) How Spotify Recommends Your New Favorite Artist. https://towardsdatascience.com/how-spotify-recommends-your-new-favorite-artist-8c1850512af0. Zugegriffen: 09. März 2022