INHALTSBEZOGENE SUCHE NACH GLEICHARTIGEN KLÄNGEN

Bachelorarbeit

zur
Erlangung des akademischen Grades
„Bachelor of Arts“
der Philologischen, Philosophischen und
Wirtschafts- und Verhaltenswissenschaftlichen Fakultät der
Albert-Ludwigs-Universität
Freiburg i.Br.

vorgelegt von

Elke Schächtele
aus Freiburg i. Br.

SS 2016

Musikwissenschaft

Erstgutachter
Prof. Dr. Rainer Bayreuther

Zweitgutachterin
Prof. Dr. Hannah Bast

Inhaltsverzeichnis

Zusammenfassung

In der vorliegenden Arbeit wurde ein Algorithmus für die inhaltsbezogene Suche nach gleichartigen Klängen implementiert. Eine Suchanfrage wird in Form einer Beispieldatei gestellt. Zu dieser sollen ähnlich klingende Audiodateien in einer Datenbank gefunden werden. Entscheidend für die Realisierung dieses Suchszenarios ist die Beantwortung der Frage, anhand welcher Merkmale zwei Klänge als ähnlich beurteilt werden können. Im Gegensatz zu einem Großteil der bisherigen Forschungsarbeiten wird in dieser Arbeit eine umfangreichen Literaturübersicht über verschiedene Forschungsperspektiven auf das Phänomen Klang und dessen Wahrnehmung durch den Menschen der Implementation des Algorithmus voran gestellt. Der Algorithmus wird anhand eines Benchmarks evaluiert, der mithilfe von empirischen Daten erstellt wurde. Das Resultat legt nahe, dass der typische Aufbau eines Algorithmus für die inhaltsbezogene Suche nach gleichartigen Klängen der menschlichen Wahrnehmung nur teilweise gerecht werden kann. Die Ergebnisse stehen im Einklang mit Erkenntnissen vergleichbarer Arbeiten: In diesen wurde in jüngster Zeit aufgrund der Komplexität des Forschungsgebiets zunehmend die Notwendigkeit des Ausbaus von interdisziplinärer Forschung betont.

1 Einführung

Durch die digitale Revolution sind Audiodaten heute in enormem Umfang für jeden verfügbar [62, 93]. Aufgrund von effizienteren Download- sowie Speichermöglichkeiten werden bereits private Musiksammlungen schnell so groß, dass es für den Besitzer schwer ist, den Überblick zu behalten. Noch größere Datenmengen entstehen auf Internetplattformen, bei denen regelmäßig Audiodaten von Nutzern hochgeladen werden können. Ein populäres Beispiel ist der Online-Musikdienst SoundCloud.[1] Hier wurden nach Angaben der Betreiber bereits 2012 jede Minute Audiodaten mit einer Gesamtdauer von über zehn Stunden hochgeladen und mit diesen über 180 Millionen Nutzer erreicht [100]. Obwohl Musik einen Großteil der Audiodaten im Internet ausmacht, gibt es auch zahlreiche Klangarchive – zu viele, um an dieser Stelle einen vollständigen Überblick leisten zu können. Dazu gehören beispielsweise allein über 50 nicht kommerzielle Klangarchive [107] oder knapp 100 sogenannte Sound Maps, bei denen Klänge mittels Geodaten auf Landkarten verortet werden [68]. In den Klangarchiven sind hauptsächlich relativ kurze und isolierte Klänge gespeichert wie beispielsweise Umweltgeräusche, Soundeffekte, Gesprächsfetzen oder kleine musikalische Patterns. Zu den bekannteren gehören Soundsnap[2] mit nach eigenen Angaben 200.000 Klängen und eine Million registrieren Benutzern sowie Freesound[3] mit über 200.000 Klängen [92, Stand 2015] und 3,5 Millionen registrierten Benutzern [30, Stand 2013]. Solche Klangarchive sind beispielsweise interessant für Sound Designer, die im Bereich von virtueller Realität, Game- oder Filmindustrie tätig sein können, oder auch für Künstler und Komponisten [54, S. 1].
Mit zunehmender Anzahl an Audiodaten steigt die Notwendigkeit für effiziente Suchmechanismen. Bei der Suche nach einer Audiodatei in einer Datenbank bieten sich grundsätzlich zwei Ansatzmöglichkeiten: Die textbasierte und die inhaltsbezogene Suche [75, S. 355–357]. Für die textbasierte Suche werden Ergebnisse über eine einfache Schlagwortsuche gefiltert. Ein Nutzer formuliert einen oder mehrere Suchbegriffe („Beethovens Fünfte“, „Klirren“, „bum bum zack“), die dann mit vorhandenen Metadaten der gespeicherten Audiodateien verglichen werden. Die Metadaten sind dabei nicht unmittelbar im Audiosignal enthalten, sondern

[1] Siehe `https://soundcloud.com/` (5. Juni 2016).
[2] Siehe `http://www.soundsnap.com/` (5. Juni 2016).
[3] Siehe `https://www.freesound.org/` (5. Juni 2016).

müssen gemeinsam mit der Audiodatei als ergänzende textuelle Information gespeichert werden. Metadaten können sich zum Beispiel auf allgemeine Angaben zu Titel, Dauer oder Dateiformat beziehen. In vielen Fällen werden Audiodaten auch anhand einer vorgegebenen Taxonomie beispielsweise nach Genre oder Klangquelle kategorisiert. Erweiterte Möglichkeiten bieten das sogenannte Social Tagging, bei dem Nutzer freie Schlagwörter (engl. tags) zu vorhandenen Audiodaten hinzufügen können, die diese im weiteren Sinne beschreiben („Lustig“, „typisch deutsch“, „Wow!<3“). Im Gegensatz zu vorher festgelegten Taxonomien entwickeln sich dadurch sogenannte Folksonomien [für typische Kategorien siehe beispielsweise 28, S. 14]. Ein Klang auf Freesound wird beispielsweise im Schnitt von 6,5 Tags beschrieben und die Folksonomie beinhaltet insgesamt knapp 56.000 Tags [Stand 2013 30, S. 412].
Die meisten Audio-Internetplattformen bieten ihren Nutzern ausschließlich die textbasierte Suchfunktion an. Diese kann allerdings aus verschiedenen Gründen problematisch sein [122, S. 1707, vgl. auch 83, S. 675]. Je größer der Datenbestand, desto aufwendiger und somit schwieriger wird es, manuell für korrekte und vollständig vorhandene Metadaten zu sorgen. Audiodateien ohne Metadaten können schlichtweg nicht gefunden werden und fehlerhafte Angaben führen zu falschen Suchergebnissen. Darüber fällt die sprachliche Beschreibung insbesondere von Klängen in der Regel schwer, vor allem, wenn nicht die Klangquelle, sondern der Klang an sich beschrieben werden soll oder kann [97, S. 192 f., 29, S. 100 ff.]. Selbst wenn ein Nutzer diese Hürde überwunden hat und seine Suchanfrage versprachlichen konnte, steht er dennoch vor dem Problem, dass jede Beschreibung zwangsläufig eine Abstraktion vom eigentlichen Gegenstand ist, die vor allem bei den freien Tags sehr subjektiv sein kann. Wenn ein Nutzer also nach einem „knarzendem“ Geräusch oder einem „traurigem“ Lied sucht, ist nicht garantiert, dass Ergebnisse mit entsprechenden Tags auch dessen Vorstellungen entsprechen.
Die Alternative zur textbasierten Suche ist die sogenannte inhaltsbezogene Suche (engl. content-based retrieval) [60, 47]. In diesem Fall bezieht sich eine Suchanfrage ausschließlich auf den Inhalt, der unmittelbar aus einem Audiosignal abgeleitet werden kann, ohne dabei den Umweg über sprachliche Beschreibungen zu gehen. Für die Realisierung dieses Szenarios muss der Nutzer in den meisten Fällen eine Beispieldatei liefern (engl. query-by-example), zu der entweder ähnliche Ergebnisse oder die zugehörigen Metadaten in der Datenbank gefunden werden sollen. Um dies zu ermöglichen, werden sogenannte Audiomerkmale aus der Beispieldatei extrahiert und mit den entsprechenden Audiomerkmalen der gespeicherten Daten verglichen.

1.1 Forschungskontext

Die inhaltsbezogene Suche von Audiodaten ist ein Teilgebiet des Forschungsbereichs Audio Information Retrieval [101, S. 27] und hat sich parallel zur zunehmenden Digitalisierung von Audiodaten Anfang der 1990er-Jahre entwickelt [50, S. 2]. Die Audio Information Retrieval Forschung lässt sich wiederum grob anhand der verwendeten Audiodaten unterteilen. Diese werden oft in die drei Kategorien Sprache, Musik und Umwelt- oder alltägliche Klänge (engl. environmental sounds) unterteilt [121, 17, 77, 87, 23], wobei letztere lange Zeit im Vergleich zu Musik und Sprache weniger im Interesse der Wissenschaft gelegen hat [19, 39, 87].
Ein zentraler Punkt bei der inhaltsbezogenen Suche ist die Bestimmung der Ähnlichkeit zwischen Klängen. Ähnlichkeit kann unterschiedlich genau aufgefasst werden [75, S. 357, 16, S. 682]. In manchen Anwendungen wird nach nahezu exakten Übereinstimmungen mit der Beispieldatei gesucht, in der Regel, um diese zu identifizieren. Man spricht in diesem Fall auch von Audio Fingerprinting [für einen guten Überblick über Audio Fingerprinting siehe 14, oder auch 42]. Ein prominentes Beispiel ist der Musik-Identifikationsdienst Shazam [115]. Dieser kann anhand kurzer Musikausschnitte entsprechende Metadaten zu Interpret, Titel und mehr liefern.
In anderen Anwendungen ist der Ähnlichkeitsbegriff hingegen weiter gefasst. Ein Beispiel ist die automatische Erstellung von Musikwiedergabelisten [61]. Hier liegt das Interesse darin, dem Nutzer neue Titel zu empfehlen, die der Musik, die er aktuell hört, in irgendeiner Weise ähnlich sind. Anhand welcher Eigenschaften die Ähnlichkeit festgemacht wird, ist abhängig von der jeweiligen Anwendung. Da Audiodaten in der Regel sehr komplex sind, bieten sich viele verschiedene Möglichkeiten für die Beurteilung von Ähnlichkeit an. Bezogen auf Musik können beispielsweise Tempo, Rhythmus, Instrumentierung oder Tonart relevante Eigenschaften sein. Bei alltäglichen Klängen ist es dagegen deutlich schwerer Merkmale zu finden, anhand derer Ähnlichkeit beurteilt werden kann. Die Definition von Ähnlichkeit sowie die Bewertung der Relevanz von Suchergebnissen werden dadurch zu komplexen Problemen.
Der Forschungsbereich innerhalb des Audio Information Retrievals, der sich auf alltägliche Klänge bezieht, ist im Vergleich zum Bereich von Musik und Sprache relativ klein. Davon konzentriert sich wiederum ein Großteil auf die Identifikation der Klangquelle einer gegebenen Beispieldatei oder, eng damit zusammenhängend, auf die automatische Klassifikation von Klängen anhand ihrer Klangquelle [für einen guten Überblick siehe 77, oder auch 17, sowie 19]. Ähnlichkeit ist zwar auch bei Forschungsgebieten wie Klassifikation oder Identifikation ein zentrales Element, allerdings mit anderem Fokus [111, S. 82]. Obwohl vergleichsweise wenige Studien den Schwerpunkt ihrer Untersuchung ausschließlich auf die inhaltsbezoge-

ne Suche nach ähnlichen Klängen legen, würde eine vollständige und detaillierte Literaturübersicht dennoch im Rahmen der vorliegenden Arbeit zu weit führen. Darüber hinaus ist es aufgrund mangelnder Standards und Reproduktionsmöglichkeiten schwer, die einzelnen Ergebnisse miteinander zu vergleichen [87, 45, 22]. Ein kurzer Überblick über allgemeine Tendenzen folgt in Abschnitt 3.1.
Nach Kenntnisstand der Autorin ermöglichen nur zwei Internetanwendungen die inhaltsbezogene Suche nach gleichartigen Klängen: Die ältere ist FindSounds[4] Die Plattform ist seit 2000 online und war nach eigenen Angaben damals die erste Internetsuchmaschine für Klänge [90]. Leider wird der zugrundeliegende Algorithmus von den Autoren nur sehr oberflächlich beschrieben.[5] Neben FindSounds ermöglicht auch Freesound seit den letzten paar Jahren die inhaltsbezogene Suche nach Klängen [30, für Details siehe Unterabschnitt 3.1.2].[6]

1.2 Problemdefinition und Vorgehensweise

In der vorliegenden Arbeit soll eine Suchmaschine für gleichartige Klänge entwickelt werden. Der Begriff Klang bezieht sich hier auf die Art von Audiodaten, wie sie in der Regel in Internet-Klangarchiven zu finden sind (für eine genauere Begriffsdefinition siehe Abschnitt 2.1). Diese entsprechen weitestgehend dem Audiotyp Umwelt- oder alltägliche Klänge. Suchanfragen sollen in Form von Beispieldateien gestellt werden können und dadurch die inhaltsbezogene Suche nach ähnlichen Dateien ermöglicht werden. Der Ähnlichkeitsbegriff ist relativ weit gefasst. Das heißt, das Interesse liegt nicht auf möglichst exakten Übereinstimmungen, wie es beispielsweise bei Shazam der Fall ist. Vielmehr soll die Suchmaschine dazu dienen, aus einer Sammlung von Klängen diejenigen herauszufiltern, die einem Klangbeispiel im weiteren Sinne ähnlich sind.
Woran sich beurteilen lässt, ob zwei Klänge ähnlich sind oder nicht, ist alles andere als offensichtlich. Daher beginnt diese Arbeit zunächst in Kapitel 2 mit einem Überblick über relevante Methoden der Klangbeschreibung, um sich ausgehend davon am Ende des Kapitels mit der Ähnlichkeit von Klängen zu befassen. Ziel ist es, geeignete Audiomerkmale zu finden, anhand derer zwei Audiosignale miteinander verglichen werden können. In Kapitel 3 wird der im Zuge dieser Arbeit implementierte Algorithmus detaillierter beschrieben. Es folgen abschließend Evaluation sowie Ausblick in Kapitel 4 und Kapitel 5.

[4] Siehe `http://www.findsounds.com/` (5. Juni 2016). Die inhaltsbezogene Suche auf FindSounds wird ausgehend von einem Klang über das Symbol gestartet.

[5] In [90] und [3] wird nur die Verwendung der Suchmaschine beschrieben, eine Demonstration des Algortihmus findet sich auf `http://www.comparisonics.com/search.html` (28. Mai 2016) sowie eine oberflächliche Beschreibung in [89].

[6] Bei Freesound wird die inhaltsbezogene Suche über das Symbol gestartet.

2 Ein Klang - Was ist das?

Um eine Suchmaschine für gleichartige Klänge entwickeln zu können, muss zunächst untersucht werden, was die Ähnlichkeit von Klängen beeinflusst. Dafür ist es unabdingbar, grundlegende Eigenschaften von Klang zu kennen. Natürlich ist die Frage nach dem Wesen eines Klanges zu weit gefasst, um sie im Rahmen dieser Arbeit vollständig beantworten zu können. Ziel dieses Kapitels ist es lediglich, ein für die vorliegenden Arbeit ausreichend gutes Verständnis davon zu vermitteln, was ein Klang ist.
Verschiedene Disziplinen setzen sich mit dem komplexen Phänomen Klang auseinander. Im Folgenden werden elementare Klangmerkmale aus der Perspektive von drei Forschungsrichtungen vorgestellt, die für diese Arbeit relevant sind: In Abschnitt 2.2 die physikalische Beschreibungsebene als allgemeine Grundlage, gefolgt von der psychoakustischen Beschreibungsebene in Abschnitt 2.3, da es um die vom Menschen wahrgenommene Ähnlichkeit von Klängen gehen soll und anschließend die technische Beschreibungsebene in Abschnitt 2.4, die die Grundlage für die Verarbeitung von Klängen in einem Rahmen wie dieser Arbeit ist. Abschließend wird in Abschnitt 2.5 zusammengefasst, wodurch verschiedene Klänge als gleichartig gelten können. Da bereits der Begriff Klang nicht eindeutig definiert ist, beginnt dieses Kapitel zunächst in Abschnitt 2.1 mit einer kurzen Begriffsdefinition.

2.1 Begriffsdefinition

Der Begriff *Klang* kommt als substantivierte Form vom Verb klingen und beschreibt im weitesten Sinne, wie sich etwas *anhört*. Da der Begriff sowohl im Alltag als auch in interdisziplinären Forschungskontext verwendet wird, kann es schnell zu Konfusionen kommen. Diese werden zusätzlich durch uneinheitliche Übersetzungen fremdsprachiger Terminologien verstärkt. Aufgrund der Vielschichtigkeit des Begriffs Klang gibt es keine allgemeingültige Begriffsdefinition. Die Begriffe Schall, Klang, Ton, Geräusch oder Klangfarbe werden teils gegensätzlich, teils synonym verwendet.
Im Rahmen der vorliegenden Arbeit gelten folgende Definitionen: Schall ist der physikalische Vorgang einer sich in Form von mechanischen Schwingungen durch ein bestimmtes Medium fortsetzenden Störung [117, S. 18]. Die Eigenschaften von

Schall werden durch akustische Größen wie Frequenz oder Amplitude beschrieben (siehe Abschnitt 2.2). Schall, der im Hörbereich des Menschen liegt, kann von diesem wahrgenommen werden. Beim Wahrnehmungsprozess werden relevante Informationen aus dem Schall gewonnen und verarbeitet und im Folgenden durch höhere kognitive Prozesse interpretiert [88, S. 4]. Man muss daher zwischen dem Schallereignis und dem aus dem Wahrnehmungsprozess resultierenden Hörereignis unterscheiden. Eigenschaften von Hörereignissen werden durch Wahrnehmungsgrößen wie Tonhöhe oder Lautheit beschrieben (siehe Abschnitt 2.3). Ein Audiosignal ist die Repräsentation einer Schallwelle in Form einer mathematischen Funktion [75, S. 57]. Mithilfe von Methoden der Signalverarbeitung können gezielt Informationen aus einem Audiosignal extrahiert werden. Diese werden auch als Audiomerkmale (engl. audio features) bezeichnet und können sich auf akustische Größen oder Wahrnehmungsgrößen beziehen (siehe Abschnitt 2.4). Die verschiedenen Perspektiven auf Klang sind schematisch in Tabelle 2.1 dargestellt.

Tabelle 2.1: Das Phänomen Klang aus verschiedenen Forschungsperspektiven

Forschungsbereich	Perspektive	Beschreibungsebene	Beispiele
Akustik	Schallereignis	Akustische Größen	Frequenz
Psychoakustik	Hörereignis	Wahrnehmungsgrößen	Tonhöhe
Signalverarbeitung	Audiosignal	Audiomerkmale	PitchYinFFT

Der Begriff Klang subsumiert im Folgenden alle im vorangegangenen Abschnitt genannten Beschreibungsebenen. In der Forschungsliteratur werden Klänge meistens in die drei Bereiche Sprache, Musik und Umwelt- oder alltägliche Klänge (engl. environmental sounds) unterteilt [77, S. 125, 46, S. 52], wobei die Unterteilung insbesondere hinsichtlich Letzterem problematisch ist [38, S. 6, 109, S. 1, für alternative Bezeichnungen siehe 38, S. 1]. In der vorliegenden Arbeit liegt der Fokus auf einer erweiterten Definition von Umweltklängen. Sie orientiert sich an der Klangvielfalt, wie sie im Alltag durch Nutzer von Social Networks, die explizit für den Austausch von Klängen und nicht von Musik oder sprachlichen Inhalten gedacht sind, hochgeladen werden. Das heißt, das Interesse liegt tendenziell auf eher kurzen Einzelklängen. Längere Aufnahmen oder Aufnahmen von akustischen Szenen mit verschiedenen Klängen werden aber auch dazu gezählt. Das gleiche gilt für kurze, musikalische Töne oder Patterns, die noch nicht als elaborierte Musikstücke gelten sowie knappe, sprachliche Phrasen oder Ausrufe. Letztere werden aber unabhängig von ihrer musikalischen Struktur oder sprachlichem Inhalt betrachtet. Diese Definition spiegelt das uneinheitliche Nutzerverhalten im Internet wider.

2.2 Das Schallereignis - Akustik

Als Teilgebiet der Physik beschäftigt sich die Akustik mit Schallwellen [43, S. 21–24]. Schallwellen breiten sich ausgehend von einer Quelle als Längswellen aus (siehe Abbildung 2.1a). Eine Stimmgabel stößt in regelmäßigen Abständen angrenzende Luftmoleküle an und versetzt diese in Schwingung. Dies führt an entsprechenden Stellen zu Erhöhung ($V+$) beziehungsweise Verringerung ($V-$) der Luftdichte. Die örtliche Störung setzt sich als Welle längs zur Schwingungsrichtung jedes einzelnen Luftteilchens fort. Die Dichteschwankungen sind in Form von Luftdruckschwankungen wahrnehmbar. Eine gängigere Darstellung von Schall ist die Abbildung der Luftdruckschwankungen in Form einer Querwelle (siehe Abbildung 2.1b). Man bezeichnet die Darstellung von Schall in Form einer Querwelle als *Wellenform* [75, S. 19–21]. Der einfachste Ursprung einer Schallwelle ist die Sinusschwingung wie sie in Abbildung 2.1b dargestellt ist. Die Sinusschwingung ist eine periodische Schwingung, das heißt, ein bestimmtes Muster wiederholt sich regelmäßig. Mathematisch ist die Sinusschwingung definiert als

$$g(t) := A \sin(2\pi(ft - \varphi)). \tag{2.1}$$

A entspricht der Amplitude (siehe Unterabschnitt 2.2.2), f der Frequenz (siehe Unterabschnitt 2.2.1) und φ der Phase einer Sinusschwingung [75, S. 42]. Die Phase beschreibt die zeitliche Verschiebung zwischen verschiedenen Sinuskurven gleicher Frequenz. Ihre Bedeutung ist für die vorliegende Arbeit vernachlässigbar, da sie keinen direkten Zusammenhang zur Wahrnehmung hat [75, S. 59].

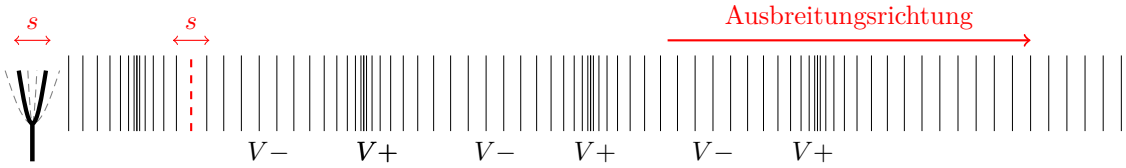
2.2.1 Frequenz

Periodische Schwingungen besitzen eine charakteristische Wellenlänge λ [43, S. 25]. Diese wird als Abstand zwischen zwei innerhalb des Schwingungsdurchlaufes gleichen Punkte gemessen. Die Wellenlängen des vom Menschen hörbaren Bereichs von Schall liegen zwischen 2 cm und 20 m. Das zeitliche Pendant zur Wellenlänge ist die Periodendauer T (siehe Abbildung 2.1b). Mithilfe der Wellenlänge λ und der Schallgeschwindigkeit c_S[1] einer Welle lässt sich ihre Frequenz f über die Formel

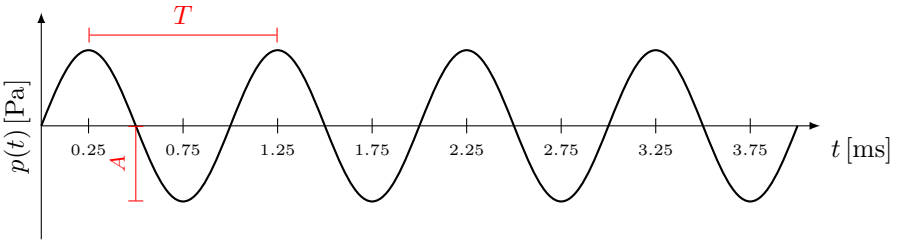
$$f = \frac{c_S}{\lambda} \tag{2.2}$$

berechnen [43, S. 35]. Die Frequenz beschreibt die Anzahl Schwingungen pro Sekunde und wird in Hertz (Hz) gemessen. Sie ist umgekehrt proportional zur Periodendauer T. Die Sinusschwingung in Abbildung 2.1b besitzt eine Periodendauer

[1] Schall breitet sich in Luft bei Raumtemperatur (20 °C) mit Geschwindigkeit $344\,\mathrm{ms}^{-1}$ aus.



(a) Längswelle



(b) Querwelle

Abbildung 2.1: Verschiedene Darstellungen von Schall.
Abbildung 2.1a: Die Erhöhung ($V+$) beziehungsweise Verringerung ($V-$) der Luftdichte breitet sich als örtliche Störung längs zur Schwingungsrichtung (s)jedes einzelnen Luftteilchens fort.
Abbildung 2.1b: Veränderung des Schalldrucks p in Pascal (Pa) an einem bestimmten Punkt in Abhängigkeit von der Zeit. Die Schwingungsrichtung ist quer zur Ausbreitungsrichtung.

von 1 ms und folglich eine Frequenz von 1.000 Hz. Die Frequenz hängt mit der wahrgenommenen Tonhöhe zusammen, wobei vereinfacht gilt: Je höher die Frequenz, desto höher der wahrgenommene Ton. Vom Menschen hörbare Frequenzen liegen im Bereich zwischen 20 Hz und 20.000 Hz.

2.2.2 Amplitude

Die physikalisch gemessene Lautstärke ist abhängig von der Amplitude A der Schallwelle [43, S. 89–91]. Die Amplitude bezieht sich hierbei auf die Sinuskurve des Schalldrucks und ist damit die maximale Druckänderung. Dies entspricht in Abbildung 2.1b dem Abstand zwischen maximaler Auslenkung und x-Achse (siehe Abbildung 2.1b). Die Intensität I von Schallwellen beschreibt die Leistung pro Flächeninhalt in Wm^{-2} und ist proportional zum Quadrat der Amplitude.
Die physikalische Lautstärke wird meist über den logarithmischen Schalldruckpegel L_P in Dezibel (dB) angegeben [43, S. 91–93]. Dieser lässt sich in Abhängigkeit

vom Schalldruck p über die Gleichung

$$L_p = 20 \log \frac{p}{p_0} dB \tag{2.3}$$

berechnen, wobei $p_0 = 2 \cdot 10^{-5}$ Pa ein festgelegter Bezugswert für die untere Hörschwelle des Menschen bei einem 1.000 Hz Sinuston ist [117, S. 28].[2]
Die Einheit Dezibel ist eine logarithmische Größe, die Dezibel-Skala eine Verhältnisskala: Eine Zunahme um n mal 10 dB bedeutet einen Intensitätsunterschied von 10^n. Der gemessene Schalldruckpegel ist auch abhängig von der Entfernung zur Schallquelle.

2.2.3 Spektrum

Isolierte Sinusschwingungen kommen als Schallwellen in der Natur nicht vor [43, S. 134–144]. Den meisten Klängen liegen komplexe Wellenformen zugrunde (siehe Abbildung 2.2 oben). Jede komplexe Wellenform kann allerdings aus einzelnen Sinuswellen zusammengesetzt gedacht werden [43, S. 146–152]. Die Frequenzen aus denen eine komplexe Schwingung zusammengesetzt ist, werden Spektrum genannt. Komplexe Schwingungen können sowohl periodisch als auch nicht-periodisch sein. Wenn die Frequenzen der einzelnen Sinuskomponenten ganzzahlige Vielfache einer gemeinsamen Grundfrequenz sind, so resultiert daraus in der Überlagerung eine periodische komplexe Wellenform (siehe Abbildung 2.2 linke Hälfte). Die Periode der resultierenden komplexen Welle entspricht der Periode der Grundschwingung. Man spricht bei solch einer periodischen Schwingung auch von harmonischer Schwingung. Bei einer harmonischen Schwingung bilden die Komponenten des Spektrums eine sogenannte Teiltonreihe. Besteht das Spektrum nicht aus so einer Teiltonreihe, resultiert daraus entweder eine teilharmonische oder eine nicht-harmonische komplexe Schwingung (siehe Abbildung Abbildung 2.2 rechte Hälfte). Die Wellenform wird neben den Frequenzen der einzelnen Komponenten auch von deren Amplituden sowie Phasenverschiebungen beeinflusst.

2.2.3.1 Fouriertransformation

Mithilfe der sogenannten Fouriertransformation kann jede komplexe Schwingung in ihre einzelnen Sinuswellen-Komponenten zerlegt werden [75, S. 39–47]. Über die

[2] Hierbei ist zu beachten, dass verschiedene Pegel berechnet werden können, die alle verkürzt als Schallpegel bezeichnet werden. Es wird unterschieden zwischen Schalldruckpegel L_P, Schallintensitätspegel L_I und Schallleistungspegel L_W [117, S. 29].

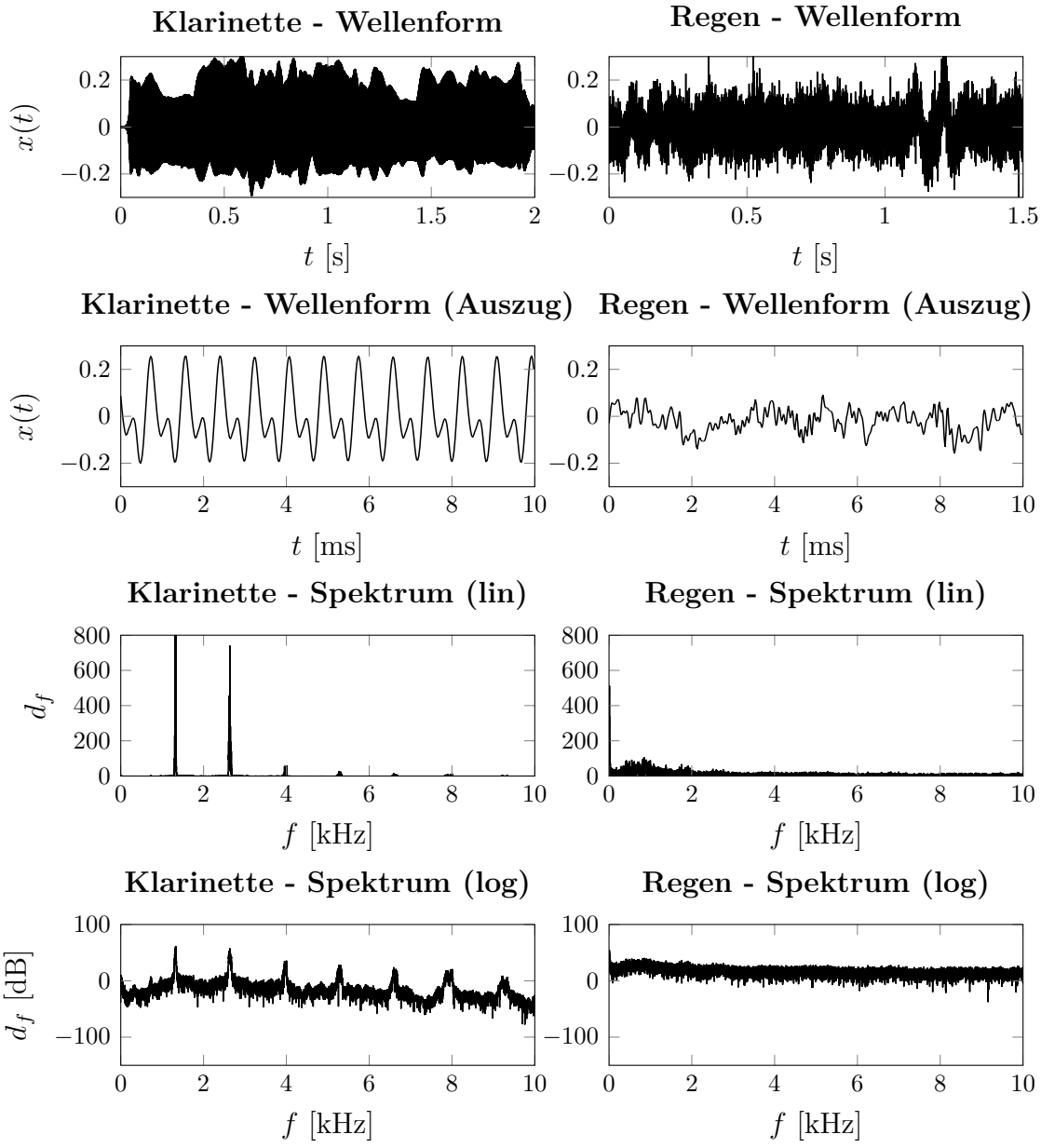


Abbildung 2.2: Wellenform (oben), Auszug (10 ms) der Wellenform (zweite Abbildung von oben) sowie Spektrum linear (zweite Abbildung von unten) und logarithmisch skaliert (unten) der Aufnahmen einer Klarinette (links) und von Regen (rechts). Aus der spektralen Darstellung wird ersichtlich, wie stark bestimmte Frequenzen im Spektrum enthalten sind: Sie zeigt den Größenkoeffizienten d_f in Abhängigkeit der Frequenz f (siehe Unterunterabschnitt 2.2.3.1). Die Klarinette evoziert eine periodische Schwingung mit klar erkennbarer Teiltonreihe mit Grundfrequenz 1.300 Hz. Regen evoziert eine nicht-periodische Schwingung und besitzt ein kontinuierliches Spektrum.

Fouriertransformation kann die zeitabhängige Darstellung von beliebigen Schallwellen (siehe Abbildung 2.2, die oberen zwei Abbildungen) in eine spektrale Darstellung (siehe Abbildung 2.2, die untere zwei Abbildungen) übertragen werden und umgekehrt. Beide Darstellungsformen haben den gleichen Informationsgehalt [117, S. 8].
Bei der Fouriertransformation wird das Signal $x(t)$ einer komplexen Welle mit Sinusschwingungen[3] $g = \cos_{f,\varphi}$ verschiedener Frequenzen f und Phasenverschiebungen φ, aber normalisierten Amplituden verglichen. Das Integral des Produkts der beiden Signale x und g dient als Maß für ihre Ähnlichkeit. Man erhält für jede Frequenz f einen Größenkoeffizienten d_f. Dieser gibt in Abhängigkeit der optimalen Phasenverschiebung φ_f Auskunft darüber, wie stark die Sinuskomponente mit Frequenz f in x enthalten ist. Mathematisch sind die beiden Koeffizienten d_f und φ_f folglich definiert als

$$d_f := \max_{\varphi \in [0,1)} \left(\int_{t \in \mathbb{R}} x(t) \cdot cos_{f,\varphi}(t) dt \right) \tag{2.4}$$

$$\varphi_f := \operatorname*{argmax}_{\varphi \in [0,1)} \left(\int_{t \in \mathbb{R}} x(t) \cdot cos_{f,\varphi}(t) dt \right). \tag{2.5}$$

Beide Koeffizienten können zusammen über den komplexen Koeffizienten c_f ausgedrückt werden. Die Fouriertransformation einer reellwertigen Funktion $x : \mathbb{R} \to \mathbb{R}$ liefert für jede Frequenz $f \in \mathbb{R}$ solch einen komplexwertigen Koeffizienten $c_f \in \mathbb{C}$. Diese können über die komplexwertige Funktion $X : \mathbb{R} \to \mathbb{C}$ mit

$$X(f) = \int_{t \in \mathbb{R}} x(t) \exp(-2\pi i f t) dt, \tag{2.6}$$

berechnet werden, wobei X die Fouriertransformierte von x ist und ihre Werte $X(f) = c_f$ die sogenannten Fourierkoeffizienten sind. Der Größenkoeffizient d_f kann über den Betrag $|X(f)|$ des Fourierkoeffizienten berechnet werden [75, S. 47]. $|X(f)|$ wird Betrags- oder Amplitudenspektrum (engl. magnitude spectrum) genannt. Da die Bedeutung der Phase in vielen Anwendungen vernachlässigbar ist [75, S. 59], ist mit Spektrum in der Regel das Betragsspektrum gemeint. Das Betragsquadrat $|X(f)|^2$ wird Leistungsspektrum (engl. power spectrum) genannt [59, S. 45].
Um die Veränderung des Spektrums über die Zeit darzustellen, kann eine sogenannte *Kurzzeit-Fouriertransformation* (engl. short-time Fourier transformation, STFT) angewendet werden [75, S. 93–94]. Bei der STFT wird eine Fouriertransformation für einzelne Klangabschnitte berechnet (auf diese Art der Verarbeitung wird in Unterabschnitt 2.4.3 genauer eingegangen). Ausgehend davon lässt sich ein

[3] Der Kosinus entspricht einer Sinusschwingung mit Phasenverschiebung $\pi/2$.

sogenanntes *Spektrogramm* anfertigen. In einem Spektrogramm wird der Größenkoeffizient d_f farblich in Abhängigkeit von Zeitpunkt t und Frequenz f dargestellt [75, S. 99]. In Abbildung 2.3 findet sich die Spektrogrammdarstellung von zehn Beispielklängen sowie eine kurze Beschreibung der einzelnen Klänge. Die zehn Klänge sind gleichzeitig die Ausgangsklänge für das zur Evaluation erstellte Datenset 1 (siehe Unterabschnitt 4.1.1). Sie können online angehört werden.[4] In Abschnitt 3.2 werden die Beispielklänge zur Visualisierung der Audiomerkmale verwendet.

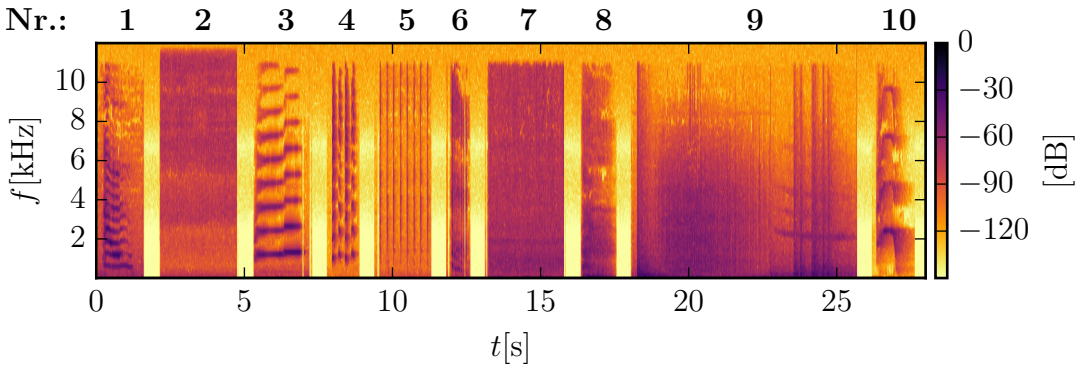
2.3 Das Hörereignis - Psychoakustik

Die physikalische Beschreibung von Schall sagt noch nichts darüber aus, wie dieser vom Menschen wahrgenommen wird. Unsere Wahrnehmung wird im Wesentlichen durch die Verarbeitung im Ohr und insbesondere im Gehirn beeinflusst. Die mechanischen Schwingungen einer Schallwelle werden bereits mit dem Einfallen durch die Ohrmuschel auf dem Weg vom Außenohr zum Innenohr aufbereitet und dort schließlich in elektrische Signale umgewandelt [27, S. 23–60, 24, S. 42–52, 116, S. 211–216]. Diese werden über den Hörnerv in das Gehirn weitergeleitet und auf verschiedenen Ebenen interpretiert, wobei diverse Neuronen auf die Verarbeitung spezifischer und unterschiedlich komplexer Klangmerkmale spezialisiert sind. Aus den verschiedenen Verarbeitungsstufen resultiert schließlich die menschliche Hörempfindung. Es ist daher wichtig, zwischen dem objektiv messbarem Schallereignis als Gegenstand der Akustik und dem subjektiv wahrgenommenen Hörereignis als Gegenstand der Psychoakustik zu unterscheiden. Grundsätzlich ist der Wahrnehmungsprozess von Klängen so kompliziert, dass die psychoakustische Froschung im Gegensatz zur Akustik noch deutliche Lücken aufweist [43, S. 402].
Die Psychoakustik versucht unter Verwendung von Methoden der Psychophysik, auditive Wahrnehmungsgrößen wie Tonhöhe oder Lautstärke in Abhängigkeit von physikalischen Stimuli zu quantifizieren [24, S. 52, 27, S. 11, zu psychoakustischen Methoden siehe 27, S. 8–15, und 105, S. 230–231]. Im Gegensatz zu physikalischen Messgrößen können konkrete Werte von Wahrnehmungsgrößen aufgrund der Subjektivität von Wahrnehmung zwischen Individuen deutlich variieren. Die Modelle zur Berechnung von Wahrnehmungsgrößen beruhen daher lediglich auf Durchschnittswerten und in der Regel gibt es mehrere Modelle, die die selbe Wahrnehmungsgröße beschreiben. [27, S. 14].
In Anlehnung an die physikalischen Parameter Frequenz, Amplitude und Spektrum ist es geradezu verlockend, drei entsprechende psychoakustische Parameter zu finden [43, S. 114]. Auf den ersten Blick könnten das die Hörempfindungen Tonhöhe,

[4] Siehe `https://raw.githubusercontent.com/ESchae/SimilarSoundSearch/master/Evaluation/D1/queries.mp3` (5. Juni 2016).

Nr.	Kurzform	Beschreibung
1	Katze	Ausgedehntes, einmaliges Miauen einer Katze
2	Hydrant	Zischen eines undichten Hydranten
3	Tür	Quietschen einer Tür mit zwei verschiedenen Tonhöhen (tief-hoch)
4	Spielzeug	Zweimaliges Quietschen eines Gummispielzeuges, jedes Quietschen besteht aus zwei Tönen nacheinander (tief-hoch)
5	Uhr	Gleichmäßiges Ticken einer Uhr
6	Lachen	Synthetisches, schnell abgespieltes Lachen
7	Regen	Gleichmäßiges Rauschen von Regen
8	Grollen	Synthetisches, tiefes und einmaliges gehauchtes Grollen
9	Feuerwerk	Feuerwerk mit mehrfachen Explosionen (eine zu Beginn und mehrere kurz nacheinander zum Ende hin), Knistern (nach der ersten Explosion) und mit hoher Tonhöhe fliegenden Raketen (gegen Ende)
10	Schrei	Sehr hoher Schrei eines Mädchens

(a) Kurzform und Beschreibung der zehn Beispielklänge



(b) Spektrogrammdarstellung der zehn Beispielklänge

Abbildung 2.3: Beschreibung und Spektrogrammdarstellung von zehn Beispielklängen. Im Vergleich zu den Spektren aus Abbildung 2.2 wurden die x-Achsen ins Vertikale gedreht. Die y-Achsen würden in einer dreidimensionalen Darstellung aus dem Blatt nach oben zeigen und werden hier farblich dargestellt: Je stärker Frequenz f zum Zeitpunkt t im Spektrum enthalten ist, desto dunkler der Farbwert. Die hellen vertikalen Balken zwischen den einzelnen Klängen sind künstlich erzeugte Stille beim Zusammenfügen der Audiodateien. Deutlich erkennbar sind harmonische Spektren als gleichmäßig horizontal übereinander angeordnete, dunkle Linien (1, 3, 4, 10) im Gegensatz zu kontinuierlichen Spektren (2, 7, 9). Gut zu erkennen sind auch Schwankungen der Tonhöhe wie beispielsweise der Tonsprung der quietschenden Tür (3). In 9 sind einzelnen Explosionen deutlich als abgetrennte vertikale Linien zu erkennen. Die horizontalen Linien bei ungefähr 24 s werden durch die mit hoher Tonhöhe fliegenden Raketen verursacht. Repetitive Strukturen sind ebenfalls klar erkennbar (4, 5).

Lautstärke und Klangfarbe sein. Diese Beobachtung reicht bis ins 19. Jahrhundert zurück, wo sie bereits von Helmholtz formulierte, einer der Pioniere der heutigen Psychoakustik [113]. Tatsächlich hängen Tonhöhe, Lautstärke und Klangfarbe jeweils im wesentlichen von Frequenz, Amplitude beziehungsweise Spektrum ab. Allerdings ist der Begriff Klangfarbe nur sehr vage definiert. Darüber hinaus sind die Zusammenhänge weder linear noch monokausal. Vielmehr werden die verschiedenen Wahrnehmungsgrößen von verschiedenen Faktoren beeinflusst. Dies soll in den folgenden Abschnitten verdeutlicht werden.

2.3.1 Tonhöhe

Die subjektive Wahrnehmung der Tonhöhe wird in Mel (mel, von engl. “melody”) gemessen [27, S. 111–118]. Es existieren verschiedene Modelle zur Berechnung der Mel-Skala [59, S. 80–81]. Der Zusammenhang zwischen Frequenz in Hz und Tonhöhe in Mel nach dem häufig verwendeten Modell von O’Shaughnessy ist in Abbildung 2.4 dargestellt. Neben der Mel-Skala existiert die sogenannte Bark-Skala zur Beschreibung der Tonhöhe.

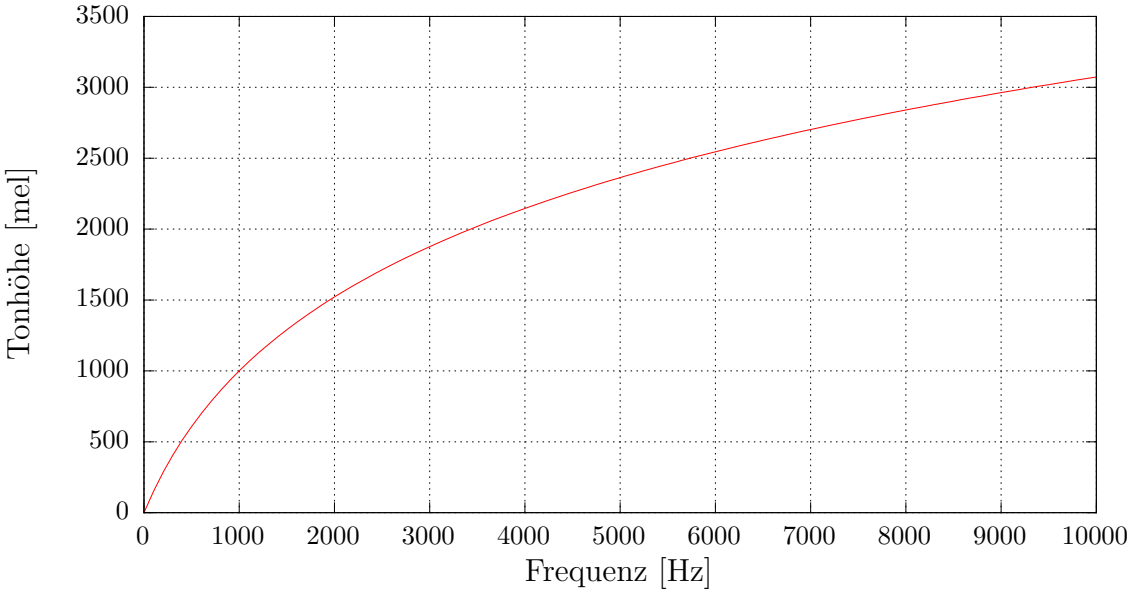


Abbildung 2.4: Zusammenhang zwischen Tonhöhe in Mel und Frequenz in Hertz. Die Tonhöhe in Mel entspricht nur bis ungefähr 500 Hz der Frequenz in Hertz. Oberhalb von 500 Hz nimmt die Frequenz im Verhältnis zur Tonhöhenwahrnehmung immer stärker zu. Dies entspricht der logarithmischen Funktionsweise des Gehörs [43, S. 113].

Bei komplexen Schwingungen wird die Tonhöhenwahrnehmung hauptsächlich durch das Spektrum beeinflusst [27, S. 119–123]. Wie in Unterabschnitt 2.2.3 beschrieben, bestehen komplexe Klänge aus verschiedenen Sinustönen, wobei ihr Spektrum entweder harmonisch, teilharmonisch oder nicht-harmonisch aufgebaut sein kann. Bei harmonischen Schwingungen verschmelzen die einzelnen Sinuskomponenten zu einer wahrnehmbaren Tonhöhe wie beispielsweise beim Klang einer Geige. Bei Grundschwingungen bis oberhalb von 1.000 Hz entspricht die wahrgenommene Tonhöhe einer Sinuswelle der Frequenz der Grundschwingung. Je tiefer allerdings die Grundschwingung, desto größer die Abweichungen zwischen der Frequenz der Grundschwingung und des wahrgenommenem Grundtons. So wird beispielsweise die Tonhöhe einer harmonischen Schwingung mit Grundschwingung von 60 Hz wie die Tonhöhe einer Sinusschwingung von 58 Hz wahrgenommen. Darüber hinaus kann es zur Wahrnehmung eines Grundtons kommen, der im Spektrum gar nicht enthalten ist. Man spricht bei diesem Phänomen von Residualton.
Ob und welche Tonhöhen wahrgenommen werden können, ist bei teilharmonischen und nicht-harmonischen Schwingungen deutlich schwerer aus dem Spektrum vorherzusagen als bei harmonischen Schwingungen. Bei teilharmonischen Schwingungen kann es zur Wahrnehmung einer konkreten Tonhöhe kommen, beispielsweise beim Klang von Trommeln. Es können aber auch verschiedene Tonhöhen wahrgenommen werden wie zum Beispiel beim Klang von Glocken [27, S. 121–122]. Bei nicht-harmonischen Schwingungen überwiegt dagegen statt einer wahrgenommenen Tonhöhe der Eindruck von Rauschen. Allerdings kann auch bei Rauschen eine Tonhöhe wahrgenommen werden [27, S. 125]. Das Maß der Tonhaltigkeit gibt auf einer Skala von undeutlich bis deutlich Auskunft darüber, inwiefern eine konkrete Tonhöhe innerhalb eines Klanges wahrgenommen wird [27, S. 135–148].
Neben Frequenz und Spektrum kann auch die Amplitude zu einem geringeren Teil Einfluss auf die Tonhöhe haben. Bei Sinuswellen kann eine Intensitätszunahme um 40 dB die Tonhöhenwahrnehmung um bis zu 3% verändern [27, S. 114]. Grundsätzlich erscheinen tiefe Töne als noch tiefer, hohe dagegen als höher. Starke Intensitätsveränderungen können allerdings maximal einen Unterschied von etwa einem Halbton bewirken [43, S. 120]. Ähnlich wie bei Sinuswellen kann die Tonhöhenwahrnehmung bei komplexen Wellen auch durch die Amplitude beeinflusst werden [27, S. 120].
Die Tonhöhenwahrnehmung kann darüber hinaus zusätzlich durch weitere Effekte beeinflusst werden. Beim Zusammenklang von mehreren Schallereignissen kann es beispielsweise zur sogenannten Maskierung, das heißt Verdeckung von Teilaspekten, Verschmelzung von Tonhöhen oder auch virtuellen Kombinationstönen kommen [zu Maskierung siehe 27, 61–74 und 114–116, zu Kombinationstönen siehe 43, S. 394–397].

2.3.2 Lautstärke

Für die wahrgenommene Lautstärke gibt es zwei psychoakustische Begriffe: Den Lautstärkepegel in Phon (phon) und die Lautheit in Sone (sone). Der Lautstärkepegel in Phon entspricht dem Lautstärkepegel in Dezibel eines 1.000 Hz-Sinustons der gleichen wahrgenommenen Lautstärke [43, S. 120–124]. Anhand der Definition wird ersichtlich, dass die wahrgenommene Lautstärke neben der Amplitude auch abhängig von der Frequenz ist. Der Zusammenhang bei Sinustönen lässt sich am sogenannten Fletcher-Munson-Diagramm veranschaulichen (siehe Abbildung 2.5).

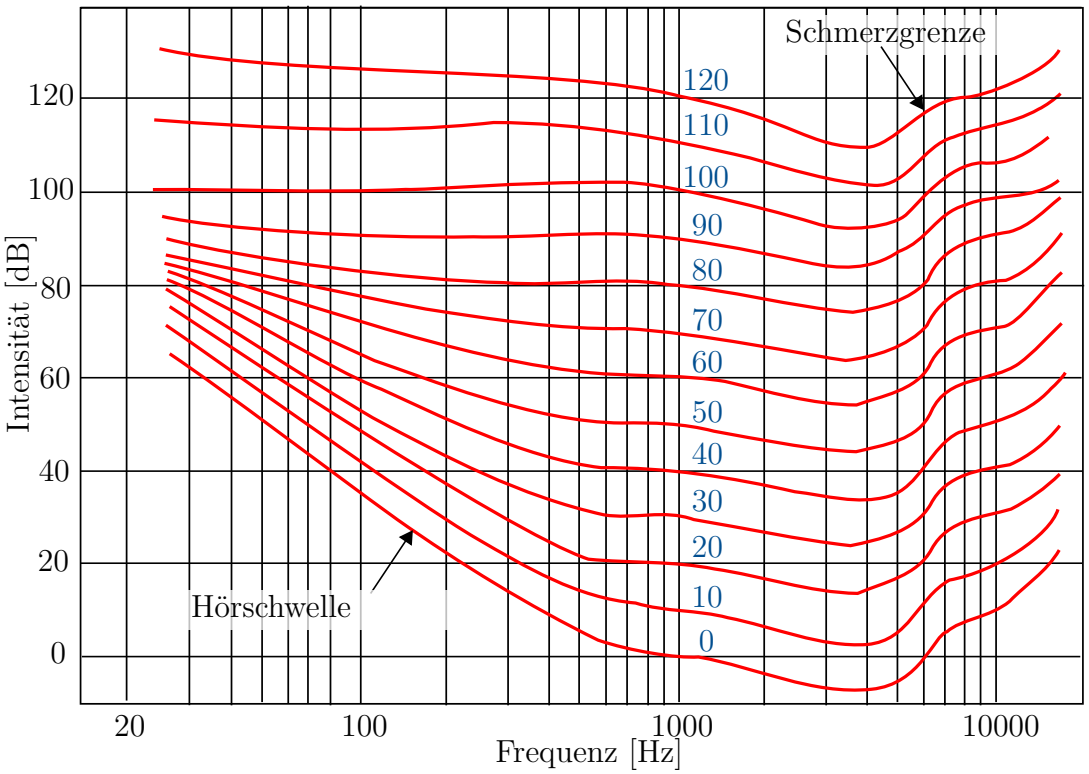


Abbildung 2.5: Fletcher-Munson-Diagramm. Die Linien sind sogenannt *Isophone*, das heißt, alle Punkte auf einer Kurve haben denselben Wert in Phon. Ein 40 Hz-Ton bei 62 dB wird also als gleich laut wahrgenommen wie ein 4.000 Hz-Ton bei ungefähr 3 dB. Gemäß der Definition entspricht bei 1.000 Hz der Wert in Phon dem Lautstärkepegel in Dezibel. Aus dem Diagramm wird ersichtlich, dass sehr tiefe und sehr hohe Frequenzen vom Menschen nicht besonders gut wahrgenommen werden. Am empfindlichsten ist das Gehör im mittleren Frequenzbereich zwischen 1.000 Hz und 6.000 Hz.[5]

Die Einheit Lautheit, gemessen in Sone (sone), ist eine dimensionslose Verhältniszahl, die Lautheitswerte miteinander vergleicht. Im Gegensatz zum Lautstärkepegel bedeutet eine Verdoppelung der Lautheit auch eine Verdoppelung der wahrgenommenen Lautstärke. Per Definition verursacht ein Sinuston mit Lautstärkepegel 40 phon die Lautheit 1 sone [43, S. 117–124].
Neben Amplitude und Frequenz können weitere Effekte die Wahrnehmung von Lautstärke beeinflussen. Bei komplexen Wellenformen ist der Wahrnehmungsprozess besonders kompliziert. Die Zusammensetzung des Spektrums spielt eine wesentliche Rolle [27, S. 208–216]. Darüber hinaus haben beispielsweise zeitliche Aspekte, Wiederholungsrate oder auch psychische Disposition Einfluss auf die wahrgenommene Lautstärke [27, S. 216–220, 29, S. 225–226].

2.3.3 Klangfarbe

Der Begriff Klangfarbe ist erst im Verlauf des 19. Jahrhunderts entstanden [76]. Seitdem wird er zwar an vielen Stellen verwendet, ist aber meistens nicht oder nur vage definiert [95]. Eine häufig angeführte Definition ist die des American National Standards Institutes [Übersetzung nach 94, S. 140]:

> Klangfarbe ist jene Eigenschaft einer Hörempfindung, nach der ein Zuhörer zwei in gleicher Weise dargebotene Schälle, die dieselbe Lautheit und dieselbe Tonhöhe hervorrufen, als unterschiedlich beurteilen kann.

Diese Definition ist aus verschiedenen Gründen problematisch [9, S. 92–93, 52, 95]: Zum einen werden nur Klänge berücksichtigt, denen eine konkrete Tonhöhe und Lautstärke zugeordnet werden kann. Dadurch sind von vorn herein viele Klänge von der Definition ausgeschlossen. Außerdem wird fälschlich die Unabhängigkeit der Parameter Klangfarbe, Tonhöhe und Lautstärke suggeriert [64, S. 45–46]. Darüber hinaus wird Klangfarbe nur negativ definiert über das, was sie nicht ist, nämlich Tonhöhe oder Lautstärke. Was Klangfarbe ist, bleibt hingegen offen. Bregman fasst daher zusammen [Übersetzung durch Autorin 9, S. 93]:

> Ich denke, die Definition von Klangfarbe der American Standards Association[6] sollte folgende sein: 'Wir wissen nicht, wie man Klangfarbe definieren soll, aber sie ist nicht Lautstärke und sie ist nicht Tonhöhe'.

[5] Die Abbildung steht gemeinfrei zur Verfügung bei der Wikimedia Foundation (`https://commons.wikimedia.org/wiki/File:4-bit-linear-PCM.svg`, 25. Mai 2016, Urheber Oarih). Lediglich die Beschriftung wurde ins Deutsche übertragen.

[6] American Standards Association war vormals Name des American National Standards Institutes

Klangfarbe ist ein Konglomerat verschiedener Wahrnehmungsgrößen [27, S. 239] und als solches eine komplexe und mehrdimensionale Größe. Es ist noch ungeklärt, wie viele und welche psychoakustischen Parameter für die eindeutige Beschreibung von Klangfarbe nötig sind. Fastl und Zwicker nennen beispielsweise Schärfe als wichtige Wahrnehmungsgröße von Klangfarbe [27, S. 239]. Die sogenannte Schärfe von Klängen, gemessen in Acum, hängt vor allem von der Hüllkurve des Spektrums sowie der zentralen Frequenz bei Klängen mit schmalem Frequenzumfang ab [27, S. 239–241]. Im Vergleich zu einem Sinuston von 1.000 Hz hat beispielsweise Rauschen mit Frequenzen ausschließlich oberhalb von 3.000 Hz eine wesentlich höhere Schärfe. Weißes Rauschen mit einem kontinuierlichen Spektrum hat im Gegensatz dazu eine geringere Schärfe, die allerdings immer noch höher als die des Sinustons ist [125, S. 149].
Ausgehend von den Zusammenhängen zwischen Frequenz und Tonhöhe sowie Amplitude und Lautstärke stellte bereits Hermann von Helmholtz 1863 mittels Ausschlussverfahren fest, dass Klangfarbe vor allem durch Spektrum und Hüllkurve eines Klangs beeinflusst werden muss [113, S. 113–114]. Mit Hüllkurve ist in diesem Fall die Veränderung der Amplitudenmaxima über die Zeit gemeint. Bis heute ist allerdings nicht umfassend geklärt, welche Aspekte von Spektrum und Hüllkurve am wichtigsten sind und wie sie sich jeweils auf die Klangfarbe auswirken [43, S. 400]. Im folgenden Abschnitt wird eine häufig angewandte Methode zur genaueren Bestimmung der Dimensionen von Klangfarbe vorgestellt.[7]

2.3.3.1 Analyse mittels Multidimensionaler Skalierung

In verschiedenen Studien wurde versucht, die Dimensionen von Klangfarbe genauer zu bestimmen und akustischen oder psychoakustischen Parametern zuzuweisen. Die meisten basieren auf dem Verfahren der Multidimensionalen Skalierung (MDS) [64, S. 36–41, 105, S. 237–240, 12, 67, 11]. Für das Verfahren werden Versuchspersonen zunächst angewiesen, zwei Klänge auf einer vorgegebenen Skala bezüglich Ähnlichkeit zu beurteilen. Jede Versuchsperson führt diese Beurteilung für alle möglichen Paare einer festen Menge von Klängen durch. Die Klänge sollten dabei so gewählt sein, dass der mögliche Einfluss von Parametern wie Tonhöhe oder Lautstärke gering gehalten wird [53]. Mittels MDS wird versucht, die Klänge in einem mehrdimensionalen Raum so darzustellen, dass die Distanzen zwischen einzelnen Klängen deren relativen Ähnlichkeiten entsprechen. Klänge mit ähnlichen Klangfarben liegen folglich näher beieinander als Klänge mit weniger ähnlichen Klangfarben. Im abschließend schwierigsten Schritt werden die einzelnen Dimensionen als Dimensionen der Wahrnehmung interpretiert. Mithilfe von Korrelationsberechnungen wird versucht, jeder Dimension einen oder mehrere

[7] Eine detailliertere Beschreibung findet sich im Anhang.

akustische oder psychoakustische Parameter in Form von Audiomerkmalen (siehe Unterabschnitt 2.4.5) zuzuweisen. Der resultierende Raum wird auch Klangfarbenraum (engl. timbre space) genannt.
MDS ist als Methode vor allem geeignet, wenn eine relativ homogene Menge an Klängen analysiert werden soll [105, S. 239]. Je heterogener die Klänge, desto wahrscheinlicher ist es, dass Ähnlichkeiten anhand von semantischen Kategorien wie beispielsweise der Klangquelle beurteilt werden [106, S. 14–15, 33, S. 174]. Darüber hinaus ist davon auszugehen, dass die Anzahl nötiger Dimensionen zunimmt, je unterschiedlicher die einzelnen Klänge sind.
Bislang haben sich ein Großteil der MDS-Studien zur Analyse von Klangfarbe auf synthetische oder instrumentale Töne konzentriert [64, S. 36–37]. Studien, die Umweltgeräusche verwendeten, beschränkten sich meistens auf einen spezifischen Bereich wie Verkehrsgeräusche, Klimaanlagen, Unterwassergeräusche oder Ähnliches [für einen Überblick siehe 105, S. 245, sowie 39, S. 840, oder auch 1, S. 64].
Aufgrund von Unterschieden bezüglich Versuchspersonen, verwendeter Klänge und MDS-Modellen, variieren die Ergebnisse zwischen den Studien unterschiedlich stark [12, S. 472]. In den meisten Fällen wurde von drei, seltener auch von zwei, vier oder fünf Dimensionen berichtet [25]. Einer dieser Dimensionen wurde in der Regel ein spektraler Parameter zugeordnet, einer weiteren dagegen ein temporaler. [66, 11, 12, 39]. Weitere Dimensionen ließen dagegen oft weniger eindeutige Interpretationen zu [siehe beispielsweise 8, S. 78]. Bislang konnten keine akustischen oder psychoakustischen Parameter zur universalen Klangfarbenbeschreibung beliebiger Klänge gefunden werden [105, S. 246]. Als wichtiges Audiomerkmale hat sich in einem Großteil der Studien jedoch der sogenannte Spectral Centroid (siehe Unterunterabschnitt 3.2.5.1) herausgestellt [105, S. 246], der mit der wahrgenommenen Schärfe von Klängen (siehe Beginn Unterabschnitt 2.3.3) zusammenhängt [59, S. 45].
Trotz der unterschiedlichen Ergebnisse gilt MDS im Gegensatz zu anderen Methoden wie beispielsweise der Analyse von sprachlichen Klangfarbenbeschreibungen [vgl. 108, 46, 112] als besonders vielversprechend [105, S. 239]. Der Klangfarbenraum sollte allerdings nie eins zu eins als zugrundeliegende kognitive Struktur, sondern stets in Abhängigkeit der verwendeten Klänge und Aufgabenstellung interpretiert werden [99, S. 35].

2.4 Das digitale Audiosignal - Signalverarbeitung

Audiosignale können sowohl in analoger als auch digitaler Form vorliegen. Die vom Menschen wahrgenommenen Schallwellen sind analoge Audiosignale [75, S. 58, 59, S. 7]. Analoge Signale sind kontinuierliche Signale. Das heißt, sowohl die zeitliche Achse als auch die Amplitude können beliebig fein abgestufte, also unendlich viele Werte annehmen. Digitale Computer können allerdings nur eine endliche Anzahl von Werten verarbeiten. Um von Computern verarbeitet werden zu können, müssen kontinuierliche Signale in sogenannte diskrete Repräsentationen umgewandelt werden. Diesen Vorgang nennt man Digitalisierung. Die Digitalisierung von Audiosignalen erfolgt in der Regel anhand von Abtastung (engl. sampling) im zeitlichen Verlauf und Quantisierung im Amplitudenverlauf.

2.4.1 Abtastung

Abtastung (engl. sampling) ist der Vorgang, durch den ein zeitkontinuierliches Signal in ein zeitdiskretes Signal übertragen wird [75, S. 60–61]. Dazu wird das zeitkontinuierliche Signal zu bestimmten Zeitpunkten abgetastet (siehe Abbildung 2.6). Das daraus resultierende zeitdiskrete Signal ist nur noch für eine endliche Menge von Zeitpunkten definiert. In der Regel ist die Abtastperiode T, das heißt der Abstand zwischen zwei Abtastpunkten (engl. samples) konstant. Man spricht in diesem Fall von äquidistanter Abtastung. Die Abtastrate oder Samplingfrequenz f_S in Hertz gibt an, wie häufig die Abtastung pro Sekunde erfolgt. Die Abtastrate für CD Aufnahmen beträgt beispielsweise 44.100 Hz.
Obwohl bei der Abtastung Informationen des Originalsignals verloren gehen, kann dieses weitestgehend rekonstruiert werden, sofern das sogenannte Abtasttheorem eingehalten wurde. Dieses besagt, dass die Abtastrate mindestens doppelt so hoch wie die höchste im Signal vorkommende Frequenz sein muss. Wird das Abtasttheorem nicht eingehalten, kann es zu Störeffekten durch sogenanntes *Aliasing* kommen, was sich in Form von falsch wiedergegebenen Frequenzen äußert. Das Originalsignal bleibt hierbei erhalten, nur kommen noch zusätzliche, zumeist sehr hohe Frequenzen hinzu, die aus einem Rechenfehler entstehen.

2.4.2 Quantisierung

Die *Quantisierung* ist die Amplitudendiskretisierung eines Signals [75, S. 61–63]. Der kontinuierliche Wertebereich wird in Quantisierungsstufen unterteilt und die einzelnen Amplitudenwerte jeweils einer Quantisierungsstufe zugewiesen (siehe

Abbildung 2.6). Die Quantisierungsstufen sind alle so groß wie das Quantisierungsintervall Δ. Die Funktion, die die Amplitudenwerte den Quantisierungsstufen zuordnet, wird Quantisierer genannt.
Wie bei der Abtastung gehen auch bei der Quantisierung Informationen verloren. Bei der Quantisierung lassen sich diese allerdings nicht rekonstruieren. Der Quantisierungsfehler beschreibt die Differenz zwischen analogem und quantisiertem Amplitudenwert. Er kann verringert werden, indem das Quantisierungsintervall Δ verkleinert wird, was wiederum zu mehr Quantisierungsstufen führt. Je größer die Anzahl an Quantisierungsstufen, desto mehr Bits werden benötigt, um diese digital zu repräsentieren. Die Wortbreite w mit $w = \log_2 M$ gibt an, wie viele Bits für die Repräsentation nötig sind, wobei M die Anzahl der Quantisierungsstufen ist [59, S. 11]. Die Wortbreite bei CD Aufnahmen beträgt beispielsweise in der Regel 16 Bits. Entsprechend sind $2^{16} = 65536$ Quantisierungsstufen möglich.

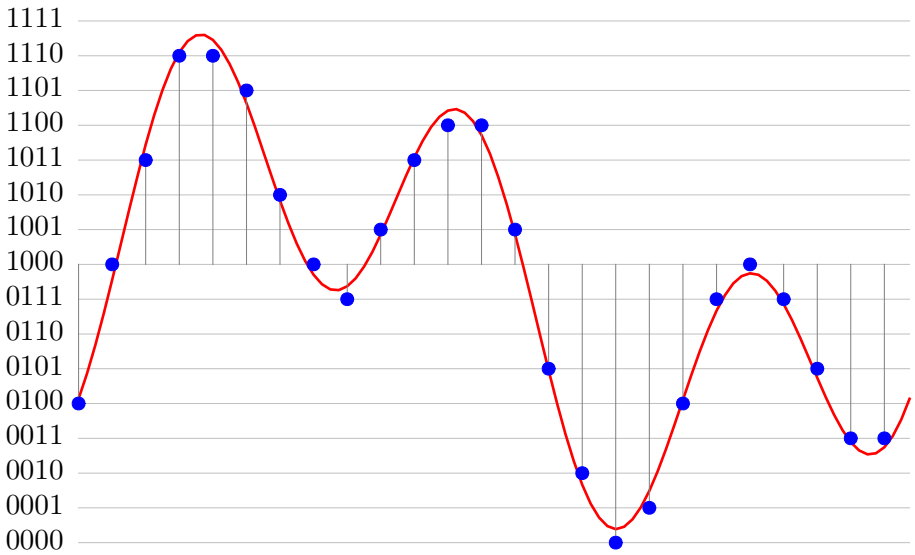


Abbildung 2.6: Digitalisierung eines analogen Signals. Das kontinuierliche Signal (rot) wird zu bestimmten Zeitpunkten abgetastet (blau). Angenommen, die Dauer wäre 1 s, entspräche dies einer Abtastrate von 25 Hz. Jeder Abtastwert wird einer der 16 Quantisierungsstufen zugeteilt. Für die Darstellung der 16 Quantisierungsstufen genügen vier Bits (links). Angenommen, die Bits repräsentieren ganze Zahlen beginnend bei Null, könnte das digitale Audiosignal als Vektor $s = [4, 8, 11, 14, 14, 13, 10, ..., 3, 3]$ dargestellt werden.[8]

[8] Die Abbildung wurde mit Abweichungen nach Vorlage aus der Wikimedia Foundation (`https://commons.wikimedia.org/wiki/File:4-bit-linear-PCM.svg`, 24. Mai 2016, Urheber Aquegg) erstellt.

2.4.3 Framebasierte Verarbeitung

Um Berechnungen auf dem Audiosignal durchzuführen, wird dieses häufig in einzelne Abschnitte (engl. frame oder auch block) der Länge $\mathcal{K}$ in Samples unterteilt. Jeder dieser *Frames* wird dann nacheinander verarbeitet. Die *framebasierte Verarbeitung* besitzt einerseits rein technischen Nutzen wie reduzierte Speicherzuweisung [59, S. 19]. Andererseits kann dadurch der zeitlichen Veränderung von Klängen Rechnung getragen werden. In der Regel wird das Audiosignal so unterteilt, dass sich die einzelnen Frames überlappen (siehe Abbildung 2.7).

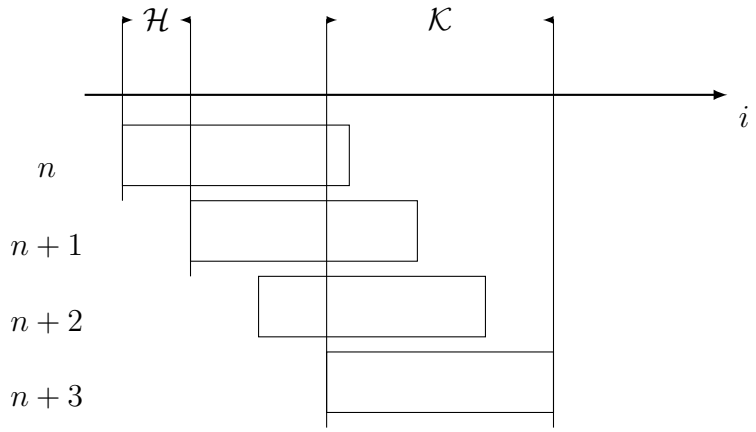


Abbildung 2.7: Schematische Darstellung der framebasierten Verarbeitung eines Signals: Das Signal mit Sampleindizes i wird in Frames der Länge $\mathcal{K}$ und Index n unterteilt, die sich in Abhängigkeit der Hop-Size $\mathcal{H}$ überlappen [aus 59, S. 19]

Der Abstand zwischen dem Beginn zweier benachbarter Frames ist die sogenannte *Hop-Size* $\mathcal{H}$ in Samples. Die Indizes der Samples $i_s(n)$ am Anfang und $i_e(n)$ am Ende eines Frames n sind folglich gegeben als [59, S. 19]

$$i_s(n) = i_s(n-1) + \mathcal{H}\,, \tag{2.7}$$

$$i_e(n) = i_s(n) + \mathcal{K} - 1\,. \tag{2.8}$$

2.4.4 Schnelle Fouriertransformation

Die Fouriertransformation wie in Unterunterabschnitt 2.2.3.1 beschrieben kann auf diskrete Signale übertragen werden.[9] Bei diskreten Signalen ist auch das Spektrum

[9] Die mathematische Übertragung auf diskrete Signale würde an dieser Stelle zu weit führen, siehe dazu [59, S. 192-196].

diskret. Einzelne Frequenzen werden zu diskreten Frequenzklassen (engl. frequency bins) k zusammengefasst [59, S. 195]. Allgegenwärtig in der Verarbeitung digitaler Audiosignale ist die STFT (siehe Unterunterabschnitt 2.2.3.1) für diskrete Signale [59, S. 20-21]. Die STFT liefert einen Koeffizienten $X(k, n)$. Dieser bezeichnet den Fourierkoeffizienten für Frequenzklasse k und Frame n. In der Regel wird direkt mit dem Betragsspektrum gerechnet. Man erhält also für jeden Frame n einen Vektor mit Größenkoeffizienten für jede Frequenzklasse k. Aufgrund von symmetrischen Eigenschaften der diskreten Fouriertransformation reicht es bei Framelänge $\mathcal{K}$ aus, $\mathcal{K}/2$ Fourierkoeffizienten zu berechnen [59, S. 196].
Die Fouriertransformationen pro Frame können anhand der sogenannten schnellen Fouriertransformation (engl. fast Fourier transform, kurz FFT) berechnet werden. Die FFT ist eine effiziente Variante der Fouriertransformation. Im Gegensatz zur Fouriertransformation, bei der $\mathcal{O}(\mathcal{K}^2)$ Rechenoperationen nötig sind, kann die Laufzeit mithilfe der FFT auf $\mathcal{O}(\mathcal{K} \log \mathcal{K})$ reduziert werden [59, S. 197].

2.4.5 Extraktion von Audiomerkmalen

Die Menge an Rohdaten in Audiosignalen ist in der Regel zu hoch, um einzelne Audiosignale effektiv miteinander vergleichen zu können. Eine einminütige CD-Aufnahme besteht beispielsweise schon aus

$$60\,\mathrm{s} \cdot 44100 \frac{\mathrm{samples}}{\mathrm{s}} \cdot 16 \frac{\mathrm{bits}}{\mathrm{sample}} = 42.336.000\,\mathrm{bits}\,. \tag{2.9}$$

Darüber hinaus beschreiben die Rohdaten lediglich die Wellenform. Um eine kompaktere Darstellung zu erhalten, können sogenannte Audiomerkmale (engl. audio features, manchmal auch descriptors) aus dem Audiosignal extrahiert werden [59, S. 4–5]. Ein Audiomerkmal kann eine beliebige numerische Repräsentation eines Audiosignals sein [Übersetzung durch Autorin 59, S. 5]:

> Offensichtlich ist der Begriff Merkmal nicht eindeutig definiert, aber er wird für jegliche Repräsentation eines zu interpretierenden Audiosignals mit deutlich geringerer Dimensionalität verwendet.

In der Regel sind Audiomerkmale Skalare oder Vektoren. Audiomerkmale repräsentieren nur noch bestimmte Informationen des ursprünglichen Audiosignals. Die einzelnen Audiomerkmale werden in einem Merkmalsvektor zusammengefasst.
In der Forschungsliteratur wurde in den vergangenen Jahrzehnten von zahlreichen Audiomerkmalen berichtet [73, S. 72]. Die Vielfalt ist einerseits damit zu begründen, dass Audiosignale multidimensional und komplex sind. Andererseits sind sie Gegenstand verschiedener Forschungsdisziplinen, die je nach Aufgabenstellung teilweise sehr unterschiedliche Merkmale extrahieren. Es ist daher nahezu unmöglich,

einen vollständigen Überblick über alle Audiomerkmale zu erhalten. Ein sehr guter Überblick findet sich bei Mitrović, Zeppelzauer & Breiteneder [Stand 2010 73]. Die Autoren haben auf der Basis von über 200 Publikationen mehr als 70 Audiomerkmale zusammengefasst und eine umfassende Taxonomie entwickelt. Im Folgenden werden die für die vorliegende Arbeit wichtigsten Eigenschaften und Basiskomponenten von Audiomerkmalen genannt.

2.4.5.1 Eigenschaften von Audiomerkmalen

Im Kontext der vorliegenden Arbeit sind vor allem die folgenden drei Eigenschaften für die Unterscheidung von Audiomerkmalen relevant [mit leichten Anpassungen durch die Autorin in Anlehnung an 73, S. 84–87, sowie 85, S. 2903]:

Audiobereich. Audiosignale können in Abhängigkeit von verschiedenen Variablen beschrieben werden. Mögliche Darstellungen sind beispielsweise die im *Zeitbereich* (anhand der Wellenform, siehe Abschnitt 2.2) oder im *Frequenzbereich* (anhand des Spektrums, siehe Unterabschnitt 2.2.3). Audiomerkmale können anhand des Bereichs, auf den sie sich beziehen, unterschieden werden. Ein Audiomerkmal im Zeitbereich beschreibt beispielsweise Aspekte der Wellenform und einer im Frequenzbereich spektrale Eigenschaften des Audiosignals. Die meisten Audiomerkmale beziehen sich auf den Frequenzbereich [73, S. 112].

Zeitliche Auflösung. Audiomerkmale können sich bezüglich der zeitlichen Auflösung unterscheiden. *Globale* Audiomerkmale werden einmalig berechnet und liefern in der Regel einen Skalar oder Vektor, der sich auf das gesamte Signal bezieht. Die meisten Audiomerkmale sind hingegen *framebasiert* [73, S. 102]. Da framebasierte Audiomerkmale auf jedem Frame n extrahiert werden, erhält man eine Folge von Werten, die den zeitlichen Verlauf des Merkmals beschreibt. Um die Dimensionalität weiter zu reduzieren, wird diese in der Regel aggregiert (siehe Unterunterabschnitt 2.4.5.2), indem beispielsweise nur der Durchschnitt des Merkmals pro Frame gespeichert und weiterverwendet wird. Man bezeichnet diesen Ansatz als *bag-of-frames* Vorgehen. Neben dem bag-of-frames Vorgehen gibt es weitere Möglichkeiten, den zeitlichen Verlauf der Werte eines Audiomerkmals über Frames hinweg zu modellieren [dazu gehören beispielsweise Hidden Markov Models oder Gaussian Mixture Models, eine detailliertere Übersicht sowie Beschreibung der einzelnen Modelle würde an dieser Stelle zu weit führen, siehe dazu 44, oder für weitere Ansätze auch 83, S. 676–677].

Semantische Interpretation. *Wahrnehmungsbasierte* Audiomerkmale extrahieren Klangeigenschaften, denen eine Bedeutung für die menschliche Wahrnehmung zugeschrieben werden kann. Sie repräsentieren in der Regel psychoakustische Merkmale und können auf Modellen des menschlichen Gehörsinnes basieren. *Physikali-*

sche Audiomerkmale beziehen sich dagegen ausschließlich auf physikalische Merkmale und stehen dadurch nicht unmittelbar in Zusammenhang zur menschlichen Wahrnehmung.

2.4.5.2 Basiskomponenten von Audiomerkmalen

Audiomerkmale werden extrahiert, indem eine Folge von mathematischen Operationen auf das Audiosignal angewendet wird. Mitrović, Zeppelzauer & Breiteneder unterscheiden drei Basiskomponenten [73, S. 86–89]:

Transformationen überführen die numerische Repräsentation eines Audiosignals von einem Audiobereich[10] in einen anderen. Dadurch verändert sich die Interpretation des Signals. Die Fouriertransformation (Unterunterabschnitt 2.2.3.1) ist beispielsweise eine Transformation der Daten vom Zeitbereich in den Frequenzbereich.

Filter bilden eine Menge von Werten auf eine andere Menge von Werten ab. Der Audiobereich wird dabei nicht verändert. In der Regel bleibt die Anzahl der Werte dieselbe. Nimmt man den Logarithmus oder das Quadrat aller Werte, wäre dies beispielsweise ein Filter.[11]

Aggregationen sind Abbildungen einer Folge von Werten auf einen einzelnen Wert. Ein Beispiel für Aggregation ist die Berechnung statistischer Kennwerte wie Durchschnitt, Varianz, Minimum oder Maximum.

2.5 Ähnlichkeit von Klängen

In der vorliegenden Arbeit soll ein Programm implementiert werden, das auf einer technischen Ebene die Ähnlichkeit von Klängen misst und sich dabei an der durch den Menschen wahrgenommenen Ähnlichkeit von Klängen orientiert. Eine entscheidende Frage ist daher zunächst, wie Menschen die Ähnlichkeit von Klängen beurteilen. Es gibt verschiedene Ebenen von Ähnlichkeiten von Klängen. Diese werden zunächst in Unterabschnitt 2.5.1 vorgestellt. Anschließend werden Studienergebnisse zu relevanten akustischen und psychoakustischen Größen in Unterabschnitt 2.5.2 zusammengefasst. Ziel ist es, relevante psychoakustische Größen und deren Zusammenhang zu akustischen Größen ausfindig zu machen. Ausgehend davon kann dann die Auswahl der Audiomerkmale für den zu implementierenden Algorithmus getroffen werden. Dieses Kapitel endet zunächst mit

[10] Frei übersetzt nach [73, S. 86], dort *domain*.

[11] Gemäß der Definition von Mitrović, Zeppelzauer & Breiteneder. In der digitalen Signalverarbeitung ist ein Filter dagegen enger definiert als System, das nur für bestimmte Frequenzkomponenten eines Eingangssignals durchlässig ist [124, S. 823].

einem abschließendem Fazit in Unterabschnitt 2.5.3. Der implementierte Algorithmus wird dann in Kapitel 3 beschrieben.

2.5.1 Arten von Ähnlichkeit

Zwei Objekte gelten generell als ähnlich, wenn sie bestimmte gemeinsame Merkmale besitzen [13, S. 9]. Klänge können auf verschiedenen Ebenen anhand verschiedener Merkmale und deren Kombinationen verglichen werden [siehe beispielsweise 56, S. 18, 46, S. 54, oder 78, S. 98–99]. In verschiedenen Studien wurden Klänge von Versuchspersonen als ähnlich beurteilt, wenn sie akustische oder psychoakustische Wahrnehmungsgrößen teilten, von einer gleichen oder vergleichbaren Klangquelle verursacht wurden oder auch ähnliche emotionale oder kontextuale Assoziationen auslösten. Ausgehend von diesen verschiedenen Strategien unterscheiden Lemaitre et al. drei grundlegende Arten von Ähnlichkeiten: Die akustische[12], kausale und semantische Ähnlichkeit von Klängen [56, S. 18]. Inwiefern die einzelnen Ebenen unabhängig voneinander sind, ist noch nicht eindeutig geklärt. Gygi, Kidd & Watson zufolge ist davon auszugehen, dass Ähnlichkeitsbeurteilungen grundsätzlich beeinflusst werden von akustischen Eigenschaften, die für die Identifikation der Klangquelle relevant sind [39, S. 853]. Giordano, McAdams & McDonnell zeigten dagegen, dass der Fokus auf akustische Ähnlichkeiten dennoch zumindest weitestgehend möglich ist [33, S. 179].
Mehrere Studien haben gezeigt, dass Versuchspersonen Klänge bevorzugt anhand kausaler oder semantischer, und weniger anhand akustischer Merkmale kategorisieren [63, 106, 37, 36]. Eine mögliche Schlussfolgerung ist, dass die alltägliche auditive Wahrnehmung primär auf die Klangquelle ausgerichtet sein könnte. Diese veranlasste bereits 1993 Gaver zur Unterscheidung von zwei Arten des Hörens: *everyday listening* und *musical listening*, wobei ersteres vor allem auf die Bestimmung der Klangquelle und letzteres auf psychoakustische Merkmale fokussiert ist [32, eine vergleichbare Unterscheidung findet sich auch bei Schaeffer 96]. Da es in der vorliegenden Arbeit um die inhaltsbezogene Ähnlichkeit von Klängen geht, liegt das Interesse allerdings ausschließlich auf der im Sinne von Lemaitre et al. akustischen Ähnlichkeit. Die zentrale Frage ist somit, welche psychoakustischen Wahrnehmungsgrößen relevant für die akustische Ähnlichkeit von Klängen sind und wie diese aus den physikalischen Eigenschaften eines Klangs abgeleitet werden können.

[12] Die Autoren subsumieren darunter sowohl akustische als auch psychoakustische Merkmale.

2.5.2 Relevante akustische und psychoakustische Größen

Im Gegensatz zu Musik und Sprache wurde die Wahrnehmung von Umweltgeräuschen bislang in verhältnismäßig wenigen Studien untersucht [45, S. 1–2, 39, S. 839, 38, S. 1–15]. Die meisten davon untersuchen die Identifikation von Umweltklängen. Nur wenige untersuchen die akustischen oder psychoakustischen Korrelate von Ähnlichkeitsbeurteilungen von Umweltgeräuschen [39, S. 840]. Es gibt verschiedene Methoden, um relevante Merkmale zu ermitteln, die Ähnlichkeitsbeurteilungen beeinflussen [34, S. 780, 105]. Die Interpretationen der meisten Studien, die die Ähnlichkeit von Klängen untersuchen, basieren entweder auf paarweisen Vergleichen (engl. dissimilarity ratings) oder Sortieraufgaben (engl. sorting tasks). Für den ersten Fall müssen Versuchspersonen alle $N(N-1)/2$ Vergleiche zweier Klänge aus einer Menge von N Klängen auf einer Ähnlichkeitsskala beurteilen [105, S. 238]. Bei Sortieraufgaben werden sie hingegen angewiesen, aus einer Menge von Klängen anhand von Ähnlichkeiten zu kategorisieren [105, S. 241, allerdings ist das Verhältnis von Kategorisierung und Ähnlichkeit ein komplexes, siehe dazu 35]. Die Daten werden dann in den meisten Fällen mittels MDS in einem geometrischen Raum oder mittels Clusteranalyse in Form einer Baumstruktur dargestellt [105, S. 242].
Untersuchungen zu Ähnlichkeitsbeurteilungen wurden sowohl für homogene als auch heterogene Mengen von Klängen durchgeführt [33, S. 173–174]. Ein Großteil der Studien zur Wahrnehmung von Klängen beschränkte sich auf homogene Klanggruppen [für einen Überblick siehe beispielsweise 105, S. 245, 40, S. 1252]. Ihre Ergebnisse lassen daher nur Aussagen über Ähnlichkeiten innerhalb einer bestimmten Klasse von Klängen zu und sind möglicherweise nicht unmittelbar übertragbar auf so heterogene Klänge, wie sie in der vorliegenden Arbeit untersucht werden sollen. Auch die zahlreichen MDS Studien zur Klangfarbe (siehe Unterunterabschnitt 2.3.3.1) beschränkten sich meistens auf einen homogenen Bereich von Klängen. Darüber hinaus wurde in diesen der Parameter Klangfarbe unabhängig von anderen Parametern untersucht. Möchte man aber Aussagen über Ähnlichkeitsbeurteilungen von Klängen im allgemeinen treffen, darf der Einfluss von Parametern wie Lautstärke, Dauer und Tonhöhe nicht vernachlässigt werden. Susini et al. zeigten beispielsweise, dass die Lautstärke einen starken Einfluss auf die Ähnlichkeitsbeurteilungen von Klängen ähnlicher Klangfarbe hat [104]. Die Ergebnisse dieser Studien sind daher nur bedingt brauchbar für die vorliegende Arbeit. Allerdings untersuchten bislang nur wenige Studien die psychoakustischen Korrelate von Ähnlichkeitsbeurteilungen auf einer sehr heterogenen Menge von Klängen [39, S. 840]. Die drei wichtigsten sollen im Folgenden kurz zusammengefasst werden.

Bonebright 2001 verwendete 74 Klänge aus dem Alltag, die nicht bezüglich Lautstärke, Tonhöhe oder Dauer normalisiert wurden [8]. Im ersten Experiment sollten Versuchspersonen diese ausschließlich anhand von akustischen Ähnlichkeiten in frei wählbare Untergruppen sortieren. In einem zweiten Experiment mussten Versuchspersonen die 74 Klänge auf vorgegebenen Skalen bezüglich bestimmter verbal spezifizierten Klangeigenschaften (dull/sharp, unpleasant/pleasant, low/high etc.) beurteilen. Darüber hinaus wurden für alle Klänge acht Audiomerkmale extrahiert. Die Audiomerkmale repräsentierten alle niederschwellige akustische Größen. Mittels MDS wurden die Ähnlichkeitsdaten in einem dreidimensionalen Raum dargestellt. Die verbalen Klangeigenschaften und Audiomerkmale wurden in Form von Vektoren in diesem Raum abgebildet. Die erste Dimension hing mit fünf Klangeigenschaften zusammen (compact/scattered, dull/sharp, slow/fast, uninteresting/interesting, rough/smooth), die zweite mit einer (low/high), die dritte mit drei (relaxed/tense, unpleasant/pleasant, soft/loud). Die Merkmalsvektoren durchschnitten teilweise die Dimensionen, wodurch der Zusammenhang zwischen den Audiomerkmalen und den einzelnen Dimensionen nicht eindeutig zu interpretieren war. Bonebright betonte zudem, dass die verwendeten Audiomerkmale möglicherweise nicht differenziert genug waren. Auffällig war, dass Audiomerkmale, die mit Lautstärke und absoluter Dauer zusammenhängen, eine besondere Rolle spielten.

Aldrich, Hellier & Edworthy 2009 führten ein Experiment mit zehn Paaren von alltäglichen Klängen durch, die wie bei Bonebright nicht normalisiert waren [1]. Die einzelnen Paare ähnelten sich akustisch, stammten aber von verschiedenen Klangquellen (beispielsweise food frying/rain falling). Versuchspersonen mussten die paarweisen Ähnlichkeiten aller Kombinationen der insgesamt 20 Klänge auf einer zehn-Punkte-Skala beurteilen. Sie bekamen keine Anweisungen darüber, anhand welcher Merkmale die Ähnlichkeit zu beurteilen sei. Die Analyse über MDS ermittelte einen dreidimensionalen Raum. Die meisten Klangpaare lagen in diesem nah beieinander, was auch mittels Clusteranalyse bestätigt wurde. Es zeigten sich folgende signifikanten Korrelationen zu extrahierten Audiomerkmalen: Die erste Dimension korrelierte mit dem relativen Anteil von Stille innerhalb eines Klangs und unterschied kontinuierliche von unterbrochenen Klängen. Die zweite Dimension korrelierte mit der durchschnittlichen und absoluten Lautstärke. Die dritte Dimension korrelierte mit Messungen der Lautstärke in höheren Frequenzen sowie der Tonhöhe.[13]

Gygi, Kidd & Watson 2007 führten die bislang umfangreichste Studie zur Ähnlichkeit von heterogenen Klängen aus dem Alltag durch [39]. Sie erhoben die paarweisen Ähnlichkeiten von 50 alltäglichen Klängen. Diese wurden so bearbeitet,

[13] Lautstärke wurde anhand des RMS (siehe Unterunterabschnitt 3.2.3.1) berechnet, zur Tonhöhenbestimmung wurde keine Angabe gemacht.

dass die Lautstärke innerhalb der einzelnen Klänge ausgeglichen war. Jeweils zwei Klänge stammten von einer gleichen Klangquelle, waren sich aber so unähnlich wie möglich. Die Versuchspersonen mussten die paarweisen Ähnlichkeiten auf einer sieben-Punkte-Skala beurteilen. Sie wurden angewiesen, ihre Beurteilung nur auf die dargebotenen Klänge zu beziehen, wurden aber nicht auf die mögliche Unterscheidung zwischen akustischer Ähnlichkeit und Ähnlichkeit der Klangquelle hingewiesen. Mittels MDS wurden die Daten in einem dreidimensionalen Raum repräsentiert. In diesem lagen ähnliche Klänge nah beieinander, was sich auch in einer Clusteranalyse zeigte. Die Autoren extrahierten im Verhältnis zu den beiden anderen Studien deutlich mehr (> 30) Audiomerkmale. Die erste Dimension korrelierte am stärksten mit der Tonhaltigkeit (zu Tonhaltigkeit siehe dritter Abschnitt Unterabschnitt 2.3.1) der Klänge. Dies bestätigte die Beobachtung, dass anhand der ersten Dimension harmonische und nicht-harmonische Klänge geordnet wurden, da sich diese darin unterscheiden, ob eine Tonhöhe wahrgenommen wird oder nicht (siehe Unterabschnitt 2.2.3). Insgesamt konnte die erste Dimension am besten durch die Kombination von Tonhaltigkeit, Konzentration der Energieverteilung im Spektrum sowie der Stärke der Veränderungen der Hüllkurve beschrieben werden. Die zweite Dimension ordnete perkussive und kontinuierliche Klänge und ist dadurch vergleichbar mit der ersten Dimension von Aldrich, Hellier & Edworthy. Die dritte Dimension war dagegen weniger eindeutig zu interpretieren. Die Korrelationen zwischen der zweiten sowie dritten Dimension mit den akustischen Messungen waren deutlich schwächer als die der ersten Dimension. Beide konnten insgesamt weniger eindeutig anhand von Kombinationen von akustischen Messungen vorhergesagt werden.

2.5.3 Fazit

Die Schlussfolgerung aus den vorangegangenen Abschnitten ist folgende: Die Studien zur Ähnlichkeit von Klängen berichten von sehr unterschiedlichen relevanten Merkmalen und sind für die Unterstützung der Auswahl geeigneter Audiomerkmale in der vorliegenden Arbeit nur bedingt brauchbar. Einerseits gibt es zu wenige Studien, die sich konkret mit dem Thema befassen. Andererseits sind die Ergebnisse aufgrund von mangelnden Standards nur schlecht vergleichbar und nur in Abhängigkeit der verwendeten Datensets und Audiomerkmale zu interpretieren. Dennoch zeigen sich gewisse Tendenzen: Beispielsweise scheint die Unterscheidung zwischen harmonischen und nicht-harmonischen Klängen eine wichtige zu sein. Darüber hinaus scheint auch die Unterscheidung zwischen eher kurzen und unterbrochenen Klängen im Gegensatz zu kontinuierlichen Klängen eine wesentliche Rolle zu spielen. Konkrete Richtlinien für die Auswahl der Audiomerkmale für die inhaltsbezogene Suche nach gleichartigen Klängen liefern die Studienergebnisse

leider nicht.
Selbst wenn ein Klang vollständig anhand psychoakustischer Parameter beschreibbar wäre, könnten dadurch noch keine globalen Vorhersagen über die Ähnlichkeit von Klängen getroffen werden. Denn die Beurteilung von Ähnlichkeit ist komplexer als der bloße Abgleich von Parametern. Die Ähnlichkeit zweier Objekte ist dynamisch und kontextabhängig [70, S. 271]. Zwar gelten zwei Objekte als ähnlich, wenn sie bestimmte Merkmale teilen. Beim Vergleich zweier Objekte wird man allerdings immer eine beliebige Anzahl von gemeinsamen Merkmalen finden können. Ohne festen Bezugspunkt ist daher die Feststellung, dass sich zwei Objekte ähnlich sind, wenig aussagekräftig. Ein Geigenton von 440 Hz ist beispielsweise dem Klang eines Autounfalls dahingehend unähnlich, dass ersterer eine wahrnehmbare Tonhöhe hat und letzterer in der Regel nicht. Allerdings könnten beide auch als ähnlich gelten, da sie sicherlich einige gemeinsame Frequenzen in ihren Spektren aufweisen.
Es ist davon auszugehen, dass Menschen beim Vergleich von Klängen nicht alle möglichen Parameter berücksichtigen, sondern nur diejenigen, die besonders hervorstechen [65, 99, S. 34]. Welche das sind, kann je nach Klang und Aufgabenstellung variieren. So zeigte sich, dass die Ähnlichkeitsbeurteilungen beispielsweise auch beeinflusst werden können durch Aufgabenstellung [1], Expertise sowie Erkennbarkeit der Klänge [56] oder Alter [5]. Möglicherweise entscheidet sich sogar erst im Moment der Ähnlichkeitsbeurteilung, welche Merkmale den stärksten Einfluss ausüben [70, S. 275, 21, S. 65]. Darüber hinaus kann es zu interindividuellen Unterschieden kommen [12, S. 482], was innerhalb mancher MDS Modelle in Form von Spezifizitäten bereits berücksichtigt wird (siehe Unterunterabschnitt 2.3.3.1).
Ähnlichkeit ist also immer kontextabhängig und subjektiv. Dadurch ist die Ähnlichkeit von Klängen nicht eindeutig nur anhand von Audiomerkmalen beschreibbar. Insgesamt tun sich für die inhaltsbezogene Suche nach gleichartigen Klängen daher zwei grundlegende Probleme auf: Einerseits ist die ideale Menge an zu extrahierenden Merkmalen noch nicht gefunden – weder in Form von psychoakustischen Merkmalen noch in Form von Audiomerkmalen, die diese repräsentieren könnten. Andererseits kann Ähnlichket nie global berechnet werden. Dass die inhaltsbezogene Suche nach gleichartigen Klängen trotzdem möglich ist und wie mit den genannten Problemen innerhalb der vorliegenden Arbeit umgegangen wurde, wird im folgenden Kapitel detaillierter beschrieben.

3 Aufbau Algorithmus

Im Folgenden wird der im Zuge dieser Arbeit implementierte Algorithmus näher beschrieben. Es handelt sich hier lediglich um eine konzeptionelle Beschreibung. Für programmiertechnische Details sei an dieser Stelle auf die Dokumentation des Quelltextes verwiesen.[1] Für die Implementierung wurde die Programmiersprache Python verwendet. Der Aufbau entspricht dem typischen Aufbau eines inhaltsbezogenen Datenabfragesystems (engl. retrieval system) [vgl. 14, S. 276–281]; dementsprechend ist auch dieses Kapitel aufgebaut. Die grundlegende Struktur ist in Abbildung 3.1 dargestellt. Hauptkomponenten sind die drei Teile Merkmalsextraktion (siehe Abschnitt 3.2), Bildung von Merkmalsvektoren (siehe Abschnitt 3.3) und Suche (siehe Abschnitt 3.4). Aus einer Sammlung von Klängen werden jeweils die selben Audiomerkmale extrahiert und anschließend in einer Datenbank gespeichert. Für die Suche werden die Merkmale von jedem Audiosignal zu einem Merkmalsvektor zusammengefasst. Stellt ein Nutzer eine Suchanfrage in Form einer Beispieldatei, so wird aus dieser auf entsprechende Weise ein Merkmalsvektor generiert. Anschließend wird der Merkmalsvektor des Beispielklanges mit den Merkmalsvektoren der gespeicherten Klänge verglichen und eine Liste von ähnlichsten Klängen zurückgegeben. Anstatt den Merkmalsvektor der Beispieldatei mit jedem Merkmalsvektor der gespeicherten Klänge zu vergleichen, müssen bei größeren Datenbanken effiziente Suchalgorithmen verwendet werden, um die Suche in angemessener Zeit ausführen zu können.

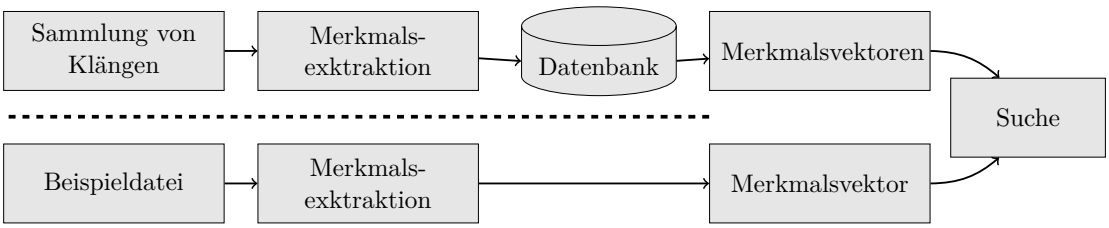


Abbildung 3.1: Aufbau des implementierten Algorithmus mit den drei wesentlichen Modulen Merkmalsextraktion, Merkmalsvektorgenerierung und Suche.

[1] Der gesamte Quellcode ist einzusehen unter `https://github.com/ESchae/SimilarSoundSearch`.

3.1 Abgrenzung zu verwandten Arbeiten

Wie bereits in der Einleitung (siehe Abschnitt 1.1) erwähnt, würde es im Rahmen dieser Arbeit zu weit führen, einen vollständigen Überblick über verwandte Arbeiten zu liefern. An dieser Stelle sollen daher hauptsächlich zwei Arbeiten näher vorgestellt werden: Zunächst SoundFisher, die erste Software zur inhaltsbezogenen Suche nach gleichartigen Klängen. Nach einem Überblick über Tendenzen von Weiterentwicklungen folgt anschließend eine kurze Beschreibung der inhaltsbezogenen Suche von Freesound.

3.1.1 Der Vorreiter: SoundFisher

Wold et al. entwickelten bereits vor zwanzig Jahren den ersten Algorithmus für inhaltsbezogene Suche nach gleichartigen Klängen [119].[2] Unter Anderem verwendeten sie diesen für die in ihrem Unternehmen Muscle Fish LLC entwickelte Software SoundFisher.[3] Diese scheint allerdings nicht mehr betrieben zu werden.
Die Autoren extrahieren pro Audiosignal auf Frames von 25 ms die vier Audiomerkmale Lautstärke, Tonhöhe, Schärfe und Bandbreite. Optional kann auch die Harmonizität extrahiert werden. Die Merkmale werden anhand von arithmetischem Mittel, Varianz (siehe Abschnitt 3.3) und Autokorrelationswerten aggregiert (bag-of-frames Vorgehen, siehe Unterunterabschnitt 2.4.5.1). Die Autokorrelation beschreibt die Ebenmäßigkeit des Verlaufs der Werte eines Merkmals [119, S. 29]. Der Merkmalsvektor besteht anschließend aus den Aggregationswerten sowie der absoluten Dauer des Audiosignals. Zur Beurteilung der Ähnlichkeit zwischen Merkmalsvektoren verwendeten die Autoren die Euklidische Distanz (siehe Unterabschnitt 3.4.1).

Nachfolgende Studien haben den Algorithmus von Wold et al. unterschiedlich erweitert. Dazu gehören beispielsweise die Ergänzung um weitere Audiomerkmale [114] oder die Entwicklung spezieller neuer Audiomerkmale [54, 103, 83]. Darüber hinaus wurden verschiedene Ansätze als Alternative zum bag-of-frames Vorgehen vorgeschlagen, zum Beispiel die Veränderung der Audiomerkmale über die Zeit zu modellieren anstatt anhand von Kennwerten zusammenzufassen [44, 26] oder das Audiosignal in unterschiedlich lange Abschnitte zu gliedern anstatt alle Frames als gleichbedeutend zu behandeln [86, 15]. In Abhängigkeit des verwendeten Modells werden auch verschiedene Distanzmaße verwendet [14, S. 280]. Manche Autoren

[2] Eine detaillierte Beschreibung findet sich im zugehörige Patent aus dem Jahr 1999 [6]. In diesem ist der Algorithmus von 1996 allerdings bereits bezüglich mancher Details erweitert. Der Abschnitt hier bezieht sich auf den ursprünglichen Algorithmus.
[3] Siehe `http://www.soundfisher.com/html/overview.html` (5. Juni 2016).

beschleunigen die Suche, indem sie die gespeicherten Klänge vorab in verschiedene Cluster aufteilen [entweder anhand grober semantischer Unterscheidung wie in 114, oder auch anhand akustischer Ähnlichkeit wie in 120].

3.1.2 Am anschaulichsten: Freesound

Font, Roma & Serra entwickelten innerhalb der Music Technology Group der Universität Pompeu Fabra das Projekt Freesound [30].[4] Dass die inhaltsbezogene Suche von Freesound an dieser Stelle genauer vorgestellt werden soll, hat mehrere Gründe: Zum einen ist sie besonders anschaulich, da sie für jeden im Internet zugänglich ist. Außerdem handelt es sich um Open Source Software.[5] Darüber hinaus bietet Freesound eine sehr gut dokumentierte Programmierschnittstelle (API). Mithilfe der API wurde das Datenset der vorliegenden Arbeit zusammengestellt und dadurch der Algorithmus von Freesound als Vergleichswert für die Evaluation verwendet (siehe Kapitel 4).
Die inhaltsbezogene Suche wird auf Freesound mithilfe von zwei innerhalb der Music Technology Group entwickelten Technologien realisiert [92, S. 2]: Für die Merkmalsextraktion wird *Essentia* verwendet, eine Open Source C++ Bibliothek für Audioanalyse und Audio Information Retrieval [7]. Die Speicherung der Merkmale und Suche wird mithilfe der Open Source C++ Bibliothek *Gaia*[6] umgesetzt. Beide Bibliotheken besitzen die Möglichkeit zur Anbindung an Python.
Zunächst werden alle 40 von Essentia bereitgestellten Audiomerkmale aus dem Bereich lowlevel extrahiert. Diese werden anhand von sechs Kennwerten aggregiert: Varianz, Durchschnitt sowie Varianz und Durchschnitt der ersten und zweiten Ableitung. Pro Audiosignal ergeben sich folglich $6 \cdot 40 = 240$ Audiomerkmale, von denen manche selbst mehrere Werte beinhalten. Um die Dimensionalität zu reduzieren, wird anschließend eine sogenannte *Hauptkomponentenanalyse* (engl. Principal Components Analysis, PCA) durchgeführt. Bei der PCA wird die Dimensionalität reduziert, indem, ausgehend von den ursprünglichen Merkmalen, neue Merkmale in Form von Linearkombinationen der extrahierten Merkmale generiert werden [59, S. 69]. Dies sind die sogenannten Hauptkomponenten. Bei Freesound werden die extrahierten Merkmale mittels PCA auf 100 Hauptkomponenten reduziert.[7] Der Merkmalsvektor besteht anschließend nur noch aus den Hauptkomponenten. Als Distanzmaß wird ebenfalls die euklidische Distanz verwendet.

[4] Freesound selbst ging 2005 online, zunächst ohne Möglichkeit zur inhaltsbezogene Suche. Die inhaltsbezogene Suche ist mindestens seit 2013 möglich [30].
[5] Quellcode einzusehen unter `https://github.com/MTG/freesound` (28. Mai 2016).
[6] Quellcode einzusehen unter `https://github.com/MTG/gaia` (28. Mai 2016).
[7] Vgl. Quellcode (`https://github.com/MTG/freesound/blob/master/similarity/similarity_settings.example.py`, 28. Mai 2016).

3.1.3 Vorliegende Arbeit

Der Algorithmus der vorliegenden Arbeit ist konzeptionell dem von Wold et al. sehr nahe. Im Gegensatz zu Freesound wurde die Auswahl der Merkmale vorab getroffen.[8] Der Algorithmus steht aber auch dem von Freesound nahe, da ebenfalls Essentia zur Merkmalsextraktion verwendet wurde. Die drei Ansätze unterscheiden sich im Wesentlichen nur durch die Wahl der Merkmale.
Ausgehend vom Fazit am Ende des vorherigen Kapitels sollten ursprünglich zwei grundlegende Neuerungen implementiert werden: Einerseits sollten die verwendeten Audiomerkmale noch stärkeren Bezug zur Wahrnehmung besitzen. Andererseits sollten Kontexteffekte stärker berücksichtigt werden. Im Verlauf der Recherche stellte sich heraus, dass beide Punkte dem aktuellen Forschungsstand nach durchaus umsetzbar wären, den Rahmen der Arbeit allerdings überschritten hätten. Eine genauere Beschreibung der Ideen findet sich daher im Ausblick in Kapitel 5. Eine deutliche Abgrenzung zu einem Großteil der verwandten Arbeiten sind allerdings die intensive Auseinandersetzung mit den akustischen und psychoakustischen Korrelaten von Ähnlichkeitsbeurteilungen von Klängen sowie die Evaluation anhand von empirischen Daten (siehe Kapitel 4).

3.2 Merkmalsextraktion

Für die Merkmalsextraktion wurde in der vorliegenden Arbeit wie bei Freesound die Bibliothek Essentia (siehe Unterabschnitt 3.1.2) verwendet, da diese eine große Auswahl an bereits implementierten Algorithmen für Audiomerkmale zur Verfügung stellt. 2015 schnitt Essentia bezüglich Recheneffizienz und Umfang an Audiomerkmalen bei einem Vergleich mehrerer Bibliotheken zur Merkmalsextraktion am besten ab [74, S. 6]. Der allgemeine Ablauf der Merkmalsextraktion ist folgender: Aus einer gegebenen Audiodatei wird über den Essentia MonoLoader[9] zunächst das Audiosignal als Vektor berechnet. Die Abtastrate liegt danach einheitlich bei 44.100 Hz. Je nach Audiomerkmal muss das Audiosignal zunächst in eine Darstellung in einem anderen Audiobereich transformiert werden. Es werden sowohl globale als auch framebasierte Audiomerkmale extrahiert. Die Framelänge $\mathcal{K}$ beträgt 2048, die Hop-Size $\mathcal{H}$ 1024 Abtastwerte. Während der Merkmalsextraktion

[8] Der Verzicht auf automatisierte Techniken der Feature Selection wie PCA ist konzeptionell begründet und hängt mit der ursprünglichen Intentionen zum Aufbau des Algorithmus zusammen, siehe Kapitel 5.

[9] Vgl. Onlinedokumentation (`http://essentia.upf.edu/documentation/reference/std_MonoLoader.html`, 28. Mai 2016).

werden die einzelnen Werte zunächst in einem sogenannten Pool gespeichert.[10] Ein Pool ist eine Datenstruktur von Essentia, die die wichtigsten Funktionen für die Merkmalsextraktion wie Speicherung und anschließende Aggregation von Audiomerkmalen bereitstellt. Sie funktioniert im wesentlichen wie ein Python Dictionary. In einem Dictionary werden Daten entsprechend einem Wörterbuch anhand von Schlüsseln mit zugehörigen Werten gespeichert (für eine schematische Darstellung des Pools während der Merkmalsextraktion siehe Abbildung 3.7).
In den folgenden Abschnitten wird zunächst die Auswahl der Merkmale begründet und anschließend die verwendeten Audiomerkmale detaillierter beschrieben. Zur Veranschaulichung wird im gesamten Abschnitt eine Zusammenstellung von zehn Klängen verwendet. Es handelt sich dabei um die Klänge aus Abbildung 2.3.

3.2.1 Auswahl der Merkmale

Es gibt zahlreiche Audiomerkmale, die aus einem Audiosignal extrahiert werden können (siehe Unterabschnitt 2.4.5) und es ist nicht bekannt, welche davon die Ähnlichkeitsbeurteilungen von Klängen am stärksten beeinflussen (siehe Abschnitt 2.5). Eine Auswahl an zu extrahierenden Audiomerkmalen zu treffen ist daher nicht einfach [73, S. 93]. Einerseits können nicht einfach alle Audiomerkmale extrahiert werden, denn der zu erstellende Merkmalsvektor soll eine möglichst kompakte Repräsentation des ursprünglichen Audiosignals sein, um die Rechenkomplexität gering zu halten. Darüber hinaus sind viele Audiomerkmale untereinander korreliert [85], das heißt, die gleichzeitige Extraktion würde zu Redundanz im Merkmalsvektor führen. Andererseits bestimmt die Auswahl der Audiomerkmale, welche Informationen des ursprünglichen Audiosignals erhalten bleiben. Im Idealfall sollen alle wichtigen Klangeigenschaften durch die Audiomerkmale repräsentiert werden. Die Qualität des Algorithmus hängt wesentlich von der Wahl der Audiomerkmale ab.
Die Auswahl der Audiomerkmale in der vorliegenden Arbeit orientierte sich hauptsächlich an folgenden Kriterien: In erster Linie wurde darauf geachtet, Audiomerkmale zu extrahieren, die eine psychoakustische Interpretation ermöglichen (zur semantischen Interpretation siehe Unterunterabschnitt 2.4.5.1). Darüber hinaus wurden Audiomerkmale ausgeschlossen, die bereichsspsezifische Aspekte von Sprache und Musik beschreiben. Es ist zwar möglich, dass auch diese zur Qualität des vorliegenden Systems beitragen könnten. Allerdings hat sich gezeigt, dass eine gute Leistungsfähigkeit von Audiomerkmalen in einem Bereich nicht automatisch auf andere Bereiche übertragbar ist [118, S. 690]. Ausgehend von dieser groben Vorauswahl wurde dann versucht, Audiomerkmale zu finden, die die wichtigsten Aspekte

[10] Vgl. Quellcode (`http://essentia.upf.edu/documentation/doxygen/classessentia_1_1Pool.html`, 28. Mai 2016).

von Klängen beschreiben und somit deren Ähnlichkeitsbeurteilung beeinflussen. Dies sind die psychoakustischen Parameter Tonhöhe (siehe Unterabschnitt 3.2.2), Lautstärke (siehe Unterabschnitt 3.2.3), Dauer (siehe Unterabschnitt 3.2.6) und Klangfarbe. Um letztere im Merkmalsvektor zu repräsentieren, wurden verschiedene Merkmale extrahiert, die sich auf Hüllkurve (siehe Unterabschnitt 3.2.4) und Spektrum (siehe Unterabschnitt 3.2.5) beziehen. Zu guter Letzt hat die Verfügbarkeit von bereits implementierten Audiomerkmalen in der Essentia Bibliothek die Auswahl im Detail beeinflusst. Wenn beispielsweise für einen konkreten Aspekt verschiedene Audiomerkmale oder Varianten in Frage gekommen wären, wurde bevorzugt das Audiomerkmal verwendet, das bereits in Essentia implementiert ist.

3.2.2 Tonhöhe: Pitch und PitchConfidence

Es gibt verschiedene Ansätze zur Berechnung der Tonhöhe aus einem Audiosignal [59, S. 91–106]. In der Regel wird versucht die Grundschwingung in einem Signal ausfindig zu machen, da diese bei harmonischen Schwingungen stark mit der wahrgenommenen Tonhöhe korreliert (siehe Unterabschnitt 2.3.1). Die Grundschwingung kann sowohl im Zeitbereich anhand der Periodizität des Audiosignals als auch im Frequenzbereich anhand der Teiltöne im Spektrum berechnet werden. Zur Repräsentation der Tonhöhe im Merkmalsvektor wurde in der vorliegenden Arbeit Essentias Algorithmus PitchYinFFT[11] verwendet. PitchYinFFT liefert für jeden Frame n die beiden Audiomerkmale Pitch (v_{PITCH}) und PitchConfidence (v_{PC}) zurück (siehe Abbildung 3.2).
Nicht alle Klänge besitzen eine konkrete Tonhöhe. Darüber hinaus kann man heutzutage relativ einfach die Tonhöhe eines beliebigen Klanges nachträglich verändern. Wer nach Klängen sucht, ist daher möglicherweise weniger an der absoluten Tonhöhe interessiert, selbst wenn diese ein wesentliches Merkmal bei der Unterscheidung von Klängen ist. Im Kontext der vorliegenden Arbeit ist daher das Maß der Tonhaltigkeit (siehe Unterabschnitt 2.3.1), modelliert anhand der PitchConfidence, ein viel wichtigeres Audiomerkmal als die konkrete Tonhöhe.

3.2.2.1 Pitch

Der Rückgabewert Pitch von PitchYinFFT ist eine Schätzung der Tonhöhe in Hz. PitchYinFFT ist eine Implementation des YinFFT Algorithmus aus [10]. YinFFT ist eine effiziente Variante des YIN Algorithmus [20], die anstatt im Zeitbereich im Frequenzbereich berechnet wird. Die Grundidee von YIN ist folgende: Gegeben sei

[11] Vgl. Onlinedokumentation (`http://essentia.upf.edu/documentation/reference/std_PitchYinFFT.html`, 28. Mai 2016).

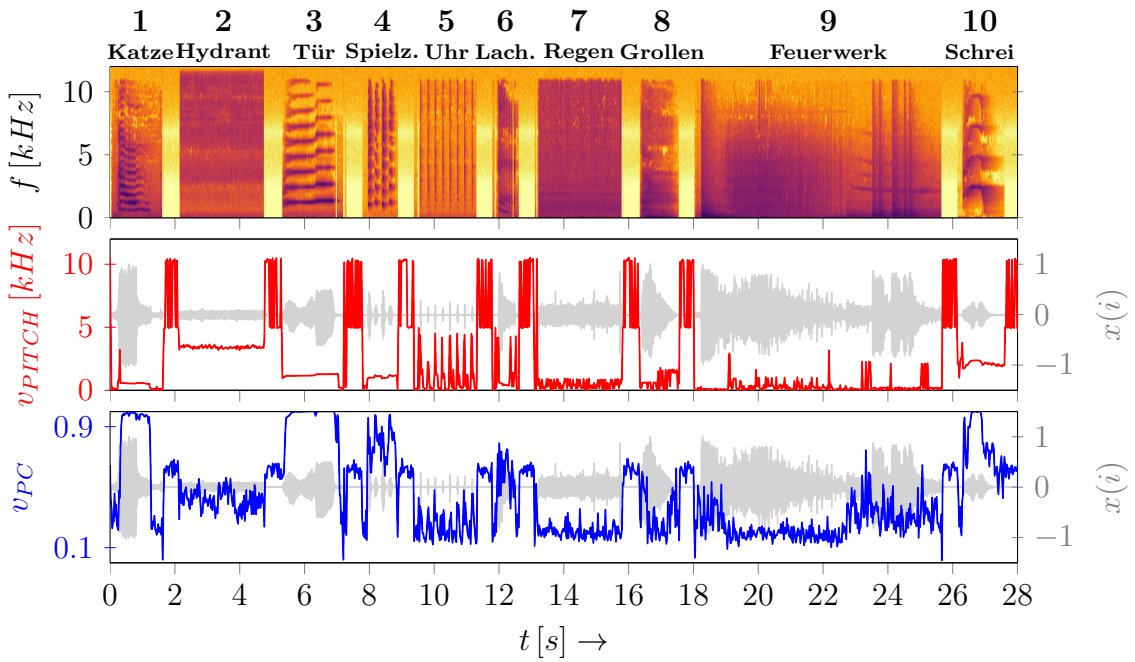


Abbildung 3.2: Spektrogramm (oben), Wellenform (mittig und unten grau), Pitch (mittig rot) sowie PitchConfidence (unten blau) der zehn Beispielklänge. Klänge mit konkreter Tonhöhe haben eine PitchConfidence nahe 1 und die jeweilige Tonhöhenschätzung entspricht dem Verlauf der Grundschwingung im Spektrum (1, 3, 4, 10). Bei Klängen, bei denen der Eindruck von Rauschen überwiegt, wird eine der tiefsten Frequenzen innerhalb der lautesten Frequenzen als Tonhöhe vorhergesagt, allerdings mit wesentlich geringerer PitchConfidence (2, 7, 9).

ein periodisches Signal $x(i)$ und eine um τ zeitverzögerte Variante $x(i+\tau)$ dieses Signals. Wenn τ genau der Periode entspricht, sind $x(i)$ und $x(i+\tau)$ identisch. Es gilt somit $x(i) - x(i+\tau) = 0$, $\forall i$, wenn τ die Periode ist. Ausgehend von der Grundidee lässt sich die pro Frame n berechnete Differenzfunktion

$$d_n(\tau) = \sum_{i=i_s(n)}^{i_e(n)} (x(i) - x(i+\tau))^2 \tag{3.1}$$

aufstellen. Da Klänge nie perfekt periodisch sind, wird die Differenzfunktion in YIN noch so erweitert, dass der Algorithmus relativ robust gegenüber Abweichungen ist. Für ein gegebenes Audiosignal wird für jeden Frame n dasjenige τ gesucht, für das die Differenzfunktion innerhalb eines vorgegebenen Schwellenwertes T minimal

wird. Dieses τ wird dann zur geschätzten Tonhöhe:

$$v_{PITCH}(n) = \arg\min_{\tau \leq T} d'_n(\tau)\,, \tag{3.2}$$

wobei d' die in YIN und YinFFT weiter modifizierte Differenzfunktion ist [für Details siehe 10, S. 77–81]. Für die Berechnung der vollständigen Differenzfunktion innerhalb eines Frames der Länge $\mathcal{K}$ wären $\mathcal{O}(\mathcal{K}^2)$ Rechenoperationen nötig [10, S. 79], da jeder Abtastwert mit jeder zeitverzögerten Variante verglichen wird. Im Frequenzbereich kann die Differenzfunktion mithilfe von zwei Fouriertransformationen berechnet werden. Wird dafür eine FFT (siehe Unterabschnitt 2.4.4) verwendet, kann die Anzahl an Rechenoperationen auf $\mathcal{O}(\mathcal{K} \log \mathcal{K})$ reduziert werden [10, S. 82].[12]

3.2.2.2 PitchConfidence

Für Klänge ohne konkrete Tonhöhe ist die geschätzte Tonhöhe nicht definiert. Daher liefert PitchYinFFT zusätzlich zur Tonhöhenschätzung noch einen Wert v_{PC} mit $0 \leq v_{PC} \leq 1$ zurück, die sogenannte PitchConfidence. Das Minimum $d'_n(\tau)$ kann als Wert für das Vertrauen, das in die Schätzung gelegt wird, verwendet werden:

$$v_{PC}(n) = \min d'_n(\tau) \tag{3.3}$$

Je niedriger, desto zuverlässiger ist die Schätzung. Die PitchConfidence von Essentias PitchYinFFT ist allerdings genau andersherum zu interpretieren: Je höher der Wert, desto zuverlässiger die Schätzung.

3.2.3 Lautstärke: Loudness und DynamicComplexity

Bei der Lautstärke gelten die gleichen Vorüberlegungen wie bei der Tonhöhe: Die absolute Lautstärke ist zwar bei Ähnlichkeitsbeurteilungen von Bedeutung, beim Suchszenario möglicherweise aber weniger wichtig, da sie leicht nachzubearbeiten ist. Daher wurde für die Repräsentation von Lautstärke im Merkmalsvektor Essentias Algorithmus DynamicComplexity[13] verwendet. Dieser ist eine Implementation eines Algorithmus von Streich [102]. Es werden zwei Werte zurückgegeben: Loudness (v_L) und DynamicComplexity (v_{DynC}). Die entsprechenden Werte für die zehn Beispielklänge sind in Abbildung 3.3 dargestellt.

[12] Man beachte, dass YIN im Gegensatz zur vollständigen Berechnung der Differenzfunktion ebenfalls effizienter ist, jedoch nicht so effizient wie die Berechnung im Frequenzbereich [10, S. 78–79, 20, S. 1922].

[13] Vgl. Onlinedolumentation (`http://essentia.upf.edu/documentation/reference/std_DynamicComplexity.html`, 28. Mai 2016).

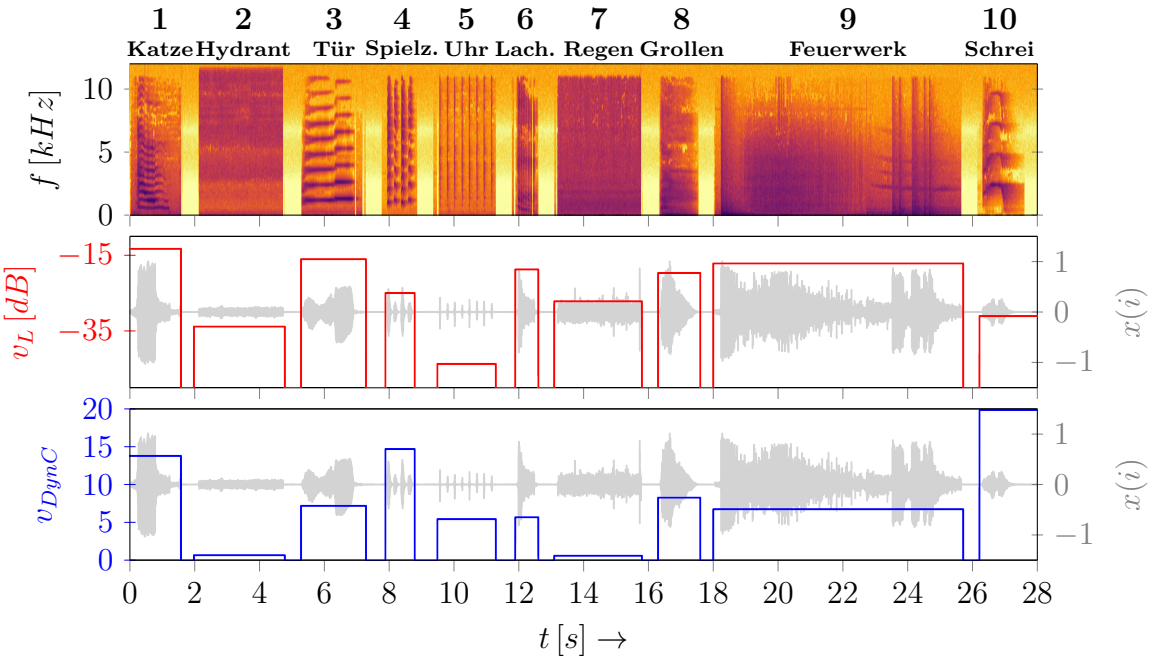


Abbildung 3.3: Spektrogramm (oben), Wellenform (mittig und unten grau), globale Lautstärke (mittig rot) sowie DynamicComplexity (unten blau) der zehn Beispielklänge. Da es sich bei den beiden Audiomerkmalen um globale Merkmale handelt, wurden sie jeweils für die einzelnen Beispielklänge berechnet und in Form einer Rechteckfunktion eingezeichnet. Für die Beurteilung der Lautstärkeschätzung sei der Leser an dieser Stelle nochmals darauf hingewiesen, dass es möglich ist, die Klänge online anzuhören. Soll man die relative Lautstärke der zehn Beispielklänge beim Anhören beurteilen, so fällt dies generell schwer und ist nicht eindeutig vom semantischen Kontext zu trennen. In jedem Fall ist 1 relativ laut im Verhältnis, was auch von der globalen Lautstärke im Schaubild entsprechend wiedergegeben wird. Die Klänge, die ausschließlich aus kontinuierlichem Rauschen bestehen, haben eine entsprechend niedrige DynamicComplexity (2, 7). Hohe DynamicComplexity besitzen Klänge mit größeren Lautstärkeschwankungen (1, 3, 4, 5). An 9 wird deutlich, dass sich die DynamicComplexity auch fälschlich neutralisieren kann, da sie lediglich als Durchschnittswert berechnet wird.

3.2.3.1 Loudness

Der Rückgabewert Loudness ist eine globale Lautstärkeschätzung. Wie bei der Tonhöhe existieren auch bei der Lautstärke verschiedene Modelle zur Berechnung [59, S. 73–78]. Die Berechnung in der Variante von Streich basiert auf einem Algorithmus von Vickers [110]. In Vickers Modell wird die Abhängigkeit der Lautstärkewahrnehmung von der Frequenz berücksichtigt. Dazu wird ein gegebenes Audiosignal mithilfe einer Gewichtungsfunktion gefiltert (zur Definition von Filter siehe Unterunterabschnitt 2.4.5.2). Die Gewichtungsfunktion ist eine grobe Näherung der Isophone im Fletcher-Munson-Diagramm (siehe Unterabschnitt 2.3.2). Aus dem vorverarbeiteten Audiosignal wird dann der sogenannte Effektivwert (engl. Root Mean Square, RMS) auf einer frame-per-frame Basis berechnet. Der RMS wird für die Berechnung der Intensität in der Literatur häufig verwendet [59, S. 73]. Er ist über

$$v_{RMS}(n) = \sqrt{\frac{1}{\mathcal{K}} \sum_{i=i_s(n)}^{i_e(n)} x(i)^2} \tag{3.4}$$

definiert als quadratisches Mittel der Amplitudenwerte pro Frame n. In Vickers' Algorithmus werden die einzelnen Werte $v_{RMS}(n)$ anschließend in $V_{dB}(n)$ in dB umgerechnet. Die globale Lautstärke v_L wird über

$$v_L = \sum_{n=n_1}^{n_\mathcal{N}} w(n) \cdot V_{dB}(n) \tag{3.5}$$

als gewichteter Durchschnitt aller Lautstärkewerte pro Frame berechnet, wobei N die Anzahl aller Frames ist. Bei Streich sind die Gewichte, ausgehend von psychoakustischen Befunden, so gewählt, dass laute Frames verhältnismäßig stärker ins Gewicht fallen [102, S. 48]. Für die Berechnung der Loudness muss im wesentlichen einmal über das gesamte Audiosignal der Länge $\mathcal{I}$ iteriert werden, um den Effektivwert des gewichteten Signals pro Frame zu berechnen, und anschließend über alle Effektivwerte, um die globale Lautstärke zu erhalten. Die Laufzeit liegt folglich im Bereich $\mathcal{O}(\mathcal{I}+\mathcal{N})$. Stille am Anfang und Ende eines Audiosignals werden in der Berechnung der globalen Lautstärke nicht berücksichtigt.

3.2.3.2 DynamicComplexity

Als DynamicComplexity definiert Streich die durchschnittliche Abweichung der Lautstärke pro Frame von der globalen Lautstärke:

$$v_{DynC} = \frac{1}{N} \sum_{n=n_1}^{n_N} |V_{dB}(n) - v_L| \ . \tag{3.6}$$

Höhere Werte von v_{DynC} entsprechen einer höheren dynamischen Komplexität. Für die Berechnung sind $\mathcal{O}(\mathcal{N})$ Rechenoperationen nötig. Die DynamicComplexity ist nur eine grobe Schätzung der Komplexität. Aspekte wie die Regelmäßigkeit oder Abruptheit von Lautstärkeschwankungen können in diesem Modell nicht reflektiert werden [102, S. 48].

3.2.4 Hüllkurve: LogAttackTime

Die LogAttackTime wird ausgehend von der zeitlichen *Hüllkurve* (engl. envelope) berechnet. Mit Hüllkurve ist in diesem Fall die Veränderung der Amplitude über die Zeit gemeint. Anhand der Hüllkurve können nicht-stationäre Klänge grob in Abschnitte wie Einschwing- (engl. attack) oder Ausklingvorgang (engl. release) unterteilen werden. Im Grunde beschreiben Audiomerkmale, die sich auf die Hüllkurve beziehen auch Aspekte, die mit der Lautstärke zusammenhängen. Sie können dennoch als eigene Gruppe von Audiomerkmalen gewertet werden, da sie auf einem anderen Audiorepräsentationsformat berechnet werden und verstärkt zeitliche Aspekte untersuchen. Die Hüllkurve an sich steht in keiner unmittelbaren Verbindung zur Wahrnehmung, da sie lediglich den Amplitudenverlauf über die Zeit beschreibt. Die Hüllkurve wird mithilfe des Essentia Envelope Algorithmus[14] aus dem Audiosignal gewonnen. Der Envelope Algortihmus ist eine Implementation des Systems aus [123, S. 110–111, ausführlicher beschrieben in 59, S. 76–77]. Für die Transformation eines Audiosignals $x(i)$ in die Darstellung als Hüllkurve $e(i)$ wird hier von jedem Abtastwert i der absolute Wert genommen und je nachdem, ob die Hüllkurve momentan ansteigt oder abfällt, unterschiedlich gewichtet. In Abbildung 3.4 ist die Hüllkurve der zehn Beispielklänge mittig in rot dargestellt. Für die Berechnung der Hüllkurve muss das Audiosignal der Länge $\mathcal{I}$ einmal iteriert werden, folglich liegt die Rechenkomplexität bei $\mathcal{O}(\mathcal{I})$.

3.2.4.1 LogAttackTime

Die LogAttackTime[15] ist ein Maß zur Beschreibung der Dauer des Einschwingvorgangs eines Klangs. Der zur Berechnung verwendete Algorithmus von Essentia ist eine Implementierung des Algorithmus aus [84, S. 9]. Start und Endpunkt des Attacks orientieren sich am Maximalwert innerhalb der Hüllkurve. Der Start

[14] Vgl. Onlinedokumentation (`http://essentia.upf.edu/documentation/reference/std_Envelope.html`, 28. Mai 2016).

[15] Vgl. Onlinedokumentation (`http://essentia.upf.edu/documentation/reference/std_LogAttackTime.html`, 28. Mai 2016).

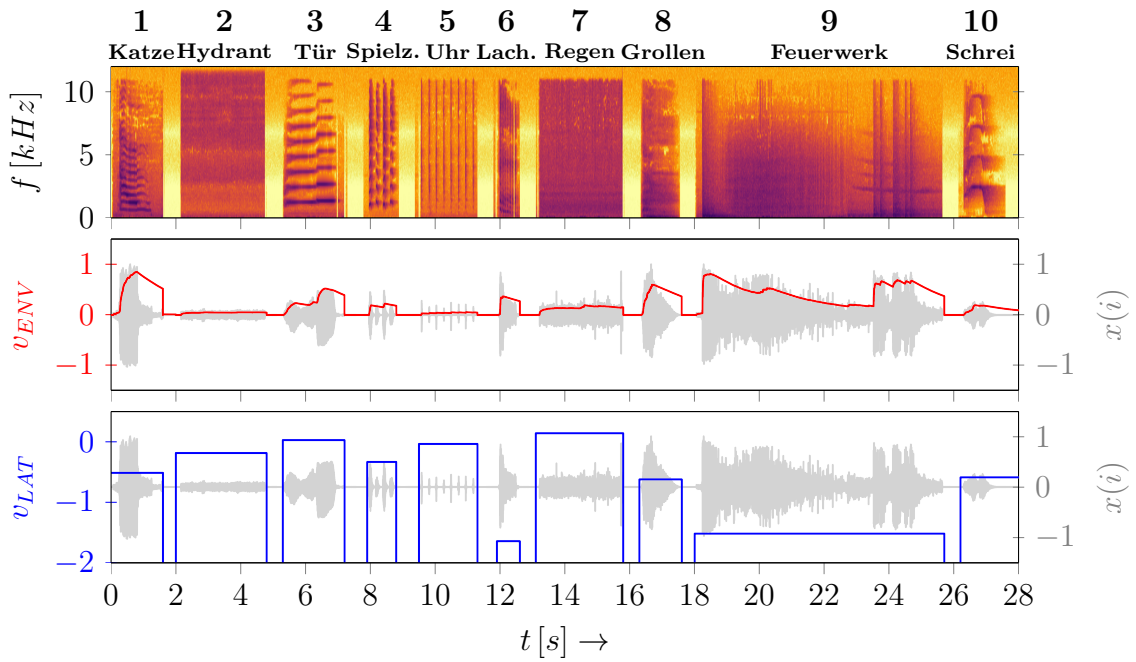


Abbildung 3.4: Spektrogramm (oben), Wellenform (mittig und unten grau), Hüllkurve (mittig rot) sowie LogAttackTime (unten blau) der zehn Beispielklänge. Die Hüllkurve wurde für die einzelnen Klänge separat berechnet. Die Dauer des Attacks der Klänge liegt im Bereich zwischen 0.02 s und 1.35 s, dementsprechend ist $-1.67 \leq v_{LAT} \leq 0.13$. Durch den Logarithmus werden die absoluten Unterschiede zwischen den einzelnen Klängen etwas relativiert. Eine niedrige LogAttackTime besitzen Klänge, die abrupt mit einem verhältnismäßig schnellen Anstieg in der Lautstärke beginnen (6, 9). In der Darstellung zeigt sich deutlich, dass die LogAttackTime für repetitive Klänge (4, 5) im vorliegenden Algorithmus weniger aussagekräftig ist, da sie global berechnet wird und nicht für jeden Einsatz. Klänge mit relativ flachem dynamischem Verlauf haben eine höhere LogAttackTime (2, 7).

des Attacks a_{start} wird definiert als Zeitpunkt, zu dem 20% des Maximalwertes überschritten werden. Dieser Schwellenwert soll die Grenze der Wahrnehmbarkeit darstellen. Das Ende des Attacks a_{end} wird definiert als Zeitpunkt, zu dem 90% des Maximalwertes überschritten werden. Die LogAttackTime ist folglich

$$v_{LAT} = \log_{10}(a_{end} - a_{start})\,. \tag{3.7}$$

Der Logarithmus der Dauer des Attacks hat sich in mehreren MDS-Studien zur Klangfarbe von musikalischen Tönen als wichtiges Audiomerkmal, das die Ähnlichkeitsbeurteilung beeinflusst, gezeigt [66, 84, S. 9]. Es ist allerdings davon auszu-

gehen, dass diese Relevanz nicht auf alle Arten von Klängen zu übertragen ist [72, S. 18]. Die LogAttackTime der zehn Beispielklänge ist in Abbildung 3.4 dargestellt. Die Rechenkomplexität der LogAttackTime liegt bei $\mathcal{O}(\mathcal{I})$.

3.2.5 Spektrum: SpectralCentroid und MFCC

Für die Berechnung der beiden folgenden Audiomerkmale wird zunächst das Audiosignal in die spektrale Darstellung überführt. Das Betragsspektrum $|X(k,n)|$ wird pro Frame n anhand von Essentias Spectrum Algorithmus[16], einer Implementierung der STFT (siehe Unterabschnitt 2.4.4) berechnet.

3.2.5.1 SpectralCentroid

Der SpectralCentroid[17] beschreibt den Schwerpunkt der Energieverteilung innerhalb des Frequenzspektrums und hängt mit der dominanten Frequenz zusammen [73, S. 116]. Er ist definiert als Verhältnis der Summe aller frequenzgewichteten Größenkoeffizienten im Spektrum zur ungewichteten Summe aller Größenkoeffizienten [59, S. 45]:

$$v_{SC}(n) = \frac{\sum_{k=0}^{\mathcal{K}/2-1} k \cdot |X(k,n)|}{\sum_{k=0}^{\mathcal{K}/2-1} |X(k,n)|} \,. \tag{3.8}$$

Die Rechenkomplexität des SpectralCentroids wird durch die Berechnung des Spektrums dominiert. Mithilfe von FFT kann dieses für einen Frame der Länge $\mathcal{K}$ in $\mathcal{O}(\mathcal{K} \log \mathcal{K})$ berechnet werden. Ausgehend vom Spektrum sind für die Berechnung des SpectralCentroids $\mathcal{O}(\mathcal{K}/2)$ Rechenoperationen nötig. Der Spectral Centroid hängt eng mit der wahrgenommenen Schärfe von Klängen zusammen und beschreibt somit einen wichtigen Aspekt von Klangfarbe. Er wurde in zahlreichen Studien als relevantes Audiomerkmal bestätigt [105, S. 234–235]. Möglicherweise liegt sein Erfolg darin begründet, dass die zugrunde liegende Klangfarbendimension verhältnismäßig konkret beschreibbar und greifbar ist.

[16] Vgl. Onlinedokumentation (`http://essentia.upf.edu/documentation/reference/std_Spectrum.html`, 6 Juni 2016).

[17] Vgl. Onlinedokumentation (`http://essentia.upf.edu/documentation/reference/std_Centroid.html`, 28. Mai 2016)

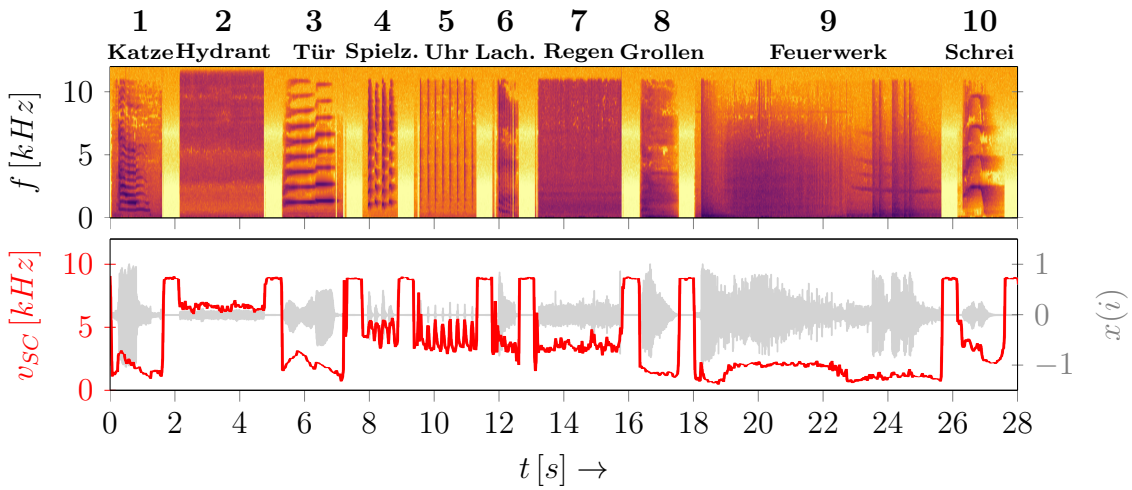


Abbildung 3.5: Spektrogramm (oben), Wellenform (unten grau) sowie Spectral Centroid (unten rot) der zehn Beispielklänge. Besonders hoch ist der Spectral Centroid bei hochfrequentem Rauschen (2). Eher niedrig ist er dagegen bei tonalen Klängen (1, 3). Der Zusammenhang mit der dominanten Frequenz ist ebenfalls gut zu erkennen.

3.2.5.2 MFCC

Wie in Unterabschnitt 2.3.3 erläutert, ist es relativ schwer, konkrete Aspekte von Klangfarbe ausfindig zu machen. Neben dem SpectralCentroid gibt es mit Sicherheit noch weitere wichtige Merkmale des Spektrums, die die Ähnlichkeit von Klängen beeinflussen. Es gibt zwar zahlreiche Audiomerkmale, die das Spektrum anhand verschiedener statistischer Merkmale beschreiben [59, S. 38–51]. Bei den meisten davon ist allerdings kein direkter Zusammenhang mit der Wahrnehmung bekannt. Ein guter Kompromiss zwischen kompakter Beschreibung des Spektrums und zumindest im weitesten Sinne auch der Klangfarbe sind die sogenannten *Mel-Frequenz-Cepstrum-Koeffizienten* (MFCC). Die MFCC wurden ursprünglich im Kontext der automatischen Spracherkennung entwickelt. Inzwischen werden sie in den meisten Anwendungen im Bereich des Audio Information Retrievals verwendet [73, S. 124]. Es gibt verschiedene Varianten der Berechnung. Im verwendeten Algorithmus von Essentia[18] werden die MFCC pro Frame n wie folgt berechnet [vgl. 73, S. 124, und 59, S. 51]: Zunächst wird das Betragsspektrum $|X(k,n)|$ berechnet. Dieses wird anschließend mithilfe entsprechender Filter auf die Mel-Skala (siehe Unterabschnitt 2.3.1) übertragen, das heißt die ursprünglich $\mathcal{K}$ Frequenzklassen

[18] Vgl. Onlinedokumentation (`http://essentia.upf.edu/documentation/reference/std_MFCC.html`, 28. Mai 2016).

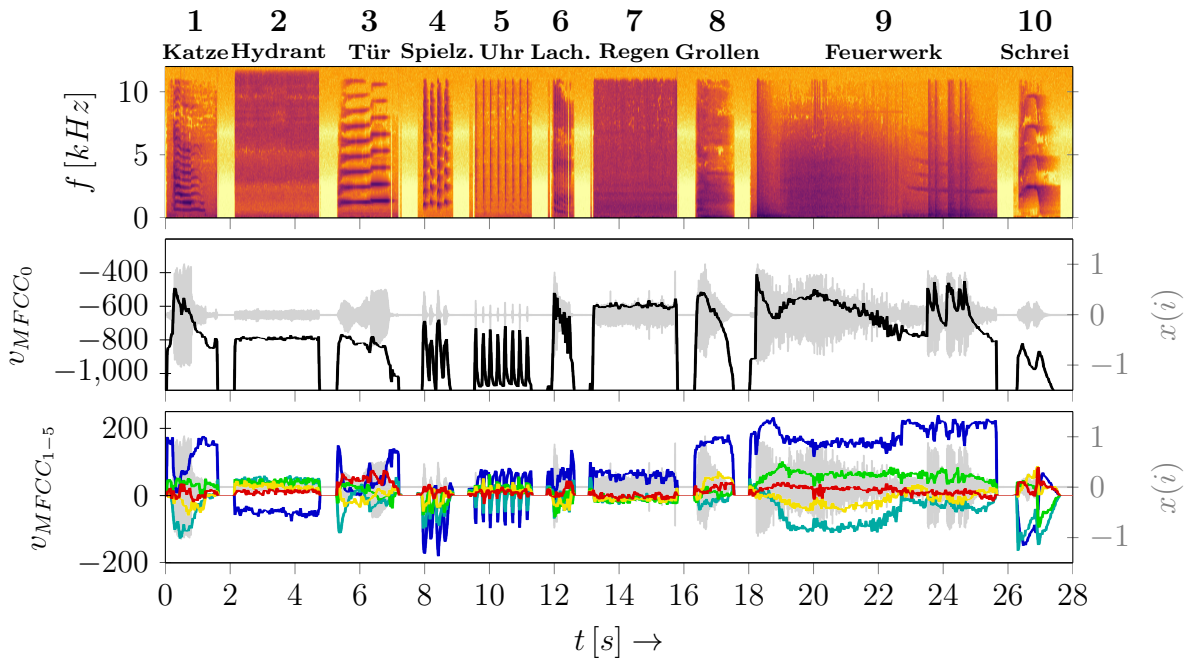


Abbildung 3.6: Spektrogramm (oben), Wellenform (mittig und unten Hintergrund) sowie die Mel-Frequenz-Koeffizienten 0 (mittig schwarz) und 1-5 (unten; blau, türkis, grün, gelb, rot in aufsteigender Reihenfolge) der zehn Beispielklänge. Deutlich erkennbar ist, dass $MFCC_0$ mit dem Amplitudenverlauf korreliert und dementsprechend keinen Bezug zur Klangfarbe hat. In der Praxis wird $MFCC_0$ daher oft ausgeschlossen. Ein unmittelbarer Zusammenhang zwischen $MFCC_{1-5}$ und Aspekten der Klangfarbe der zehn Beispielklänge ist nicht erkennbar.

des Spektrums werden auf $\mathcal{K}'$ Melbänder (engl. mel bands) zusammengefasst. In der Implementation von Essentia werden 40 Melbänder verwendet. Von dem logarithmierten Mel-Betragsspektrum $\log(|X'(k', n)|)$ wird anschließend erneut ein Spektrum berechnet. Für die Berechnung dieses Spektrums wird die sogenannte *Diskrete Kosinustransformation* (DCT) verwendet. Die DCT entspricht dem Realteil einer Fouriertransformation. Der Koeffizient j der MFCC ist formal gegeben als

$$v^j_{MFCC}(n) = \sum_{k'=1}^{K'} \log(|X'(k', n)|) \cdot \cos\left(j \cdot \left(k' - \frac{1}{2}\right) \frac{\pi}{K'}\right), \tag{3.9}$$

wobei

$$v^j_{MFCC}(n) = \sum_{k'=1}^{K'} \ldots \cdot \cos\left(j \cdot \left(k' - \frac{1}{2}\right) \frac{\pi}{K'}\right) \tag{3.10}$$

der DCT entspricht. Die Rechenkomplexität der MFCC wird durch die Berechnung des Spektrums dominiert. Mithilfe von FFT kann dieses für einen Frame der Länge $\mathcal{K}$ in $\mathcal{O}(\mathcal{K} \log \mathcal{K})$ berechnet werden. Für die Berechnung des logarithmierten Mel-Betragsspektrums sind im wesentlichen $\mathcal{O}(\mathcal{K})$ Rechenoperationen nötig, da jede Frequenzklasse k auf das zugehörige Melband k' übertragen werden muss. Da die Anzahl Melbänder $\mathcal{K}'$ konstant bei 40 liegt ist auch die Anzahl der Rechenoperationen der DFT konstant. In der Praxis werden in der Regel zwischen vier und 20 Koeffizienten verwendet. Im vorliegenden Algorithmus werden die 13 ersten Koeffizienten in den Merkmalsvektor aufgenommen.
Obwohl das Ausgangsspektrum auf die Mel-Skala übertragen wird, sind die MFCC dennoch kein wahrnehmungsbasiertes Audiomerkmal [59, S. 53]. Zwar konnte ihre Brauchbarkeit in verschiedenen Studien gezeigt werden. Der Bezug der einzelnen Koeffizienten zur Wahrnehmung ist jedoch kaum zu interpretieren. Das heißt natürlich nicht, dass solch ein unmittelbarer Bezug nicht bestehen könnte.

3.2.6 Dauer: Duration und EffectiveDuration

Für die Dauer gelten die selben Vorüberlegungen wie für die Tonhöhe und Lautstärke. Obwohl sie für die Ähnlichkeitsbeurteilung relevant ist, ist sie möglicherweise im Suchszenario weniger interessant. Im Merkmalsvektor wird die Dauer eines Klanges dennoch anhand der beiden Algorithmen Duration und EffectiveDuration aus Essentia repräsentiert.

3.2.6.1 Duration

Die Dauer eines Audiosignals in Sekunden kann einfach berechnet werden. Durch die Samplingfrequenz f_S ist festgelegt, wie viele Abtastwerte pro Sekunde im Signal enthalten sind (siehe Unterabschnitt 2.4.1). Die Dauer in Sekunden eines Audiosignals der Länge S Abtastwerte ist folglich

$$v_D = \frac{S}{f_S} \,. \tag{3.11}$$

3.2.6.2 EffectiveDuration

Die effektive Dauer v_{ED} eines Audiosignals ist nach [84, S. 11] die Dauer, in der das Signal über einem festgelegten Schwellenwert ist. Werte unterhalb des Schwellenwertes werden als unbedeutend für die Wahrnehmung beurteilt. Der Schwellenwert ist auf 40% des Maximalwertes im Audiosignal festgelegt. Die effektive Dauer gibt zwar keine Auskunft darüber, an welchen Stellen des Signals der Schwellenwert

übertreten wurde. Trotzdem erlaubt die Kombination von Dauer und effektiver Dauer eine grobe Unterscheidung von eher kurzen, schlagartigen Klängen und kontinuierlichen Klängen.

3.3 Merkmalsvektor

Der Ablauf der Generierung des Merkmalsvektors ist schematisch in Abbildung 3.7 dargestellt. Nach der Merkmalsextraktion sind im Pool (siehe Abschnitt 3.2) alle globalen und framebasierten Merkmale gespeichert. Im Falle der framebasierten Merkmale sind zunächst jeweils eine Reihe von Werten gespeichert, deren genaue Anzahl abhängig ist von der Länge des Audiosignals. Um einen Merkmalsvektor mit geringerer Dimensionalität und einheitlicher Länge für alle Audiosignale zu erhalten, werden die framebasierten Merkmale aggregiert. Die Aggregation erfolgt anhand von arithmetischem Mittelwert und Varianz [siehe 59, S. 36–37] sowie Mittelwert und Varianz der ersten Ableitung (d). Der Mittelwert ist definiert als

$$\mu_v = \frac{1}{\mathcal{N}} \sum_{n=n_1}^{\mathcal{N}} v(n). \tag{3.12}$$

Die Varianz ist ein Maß für die Streuung der Werte um den Mittelwert. Sie ist definiert als

$$\sigma_v^2 = \frac{1}{\mathcal{N}} \sum_{n=n_1}^{\mathcal{N}} (v(n) - \mu_v)^2. \tag{3.13}$$

Die Ableitung für eine diskrete Folge von Werten wird als Folge von Differenzen zwischen einem Wert und dem vorangehenden Wert berechnet. Sie ist ein Maß für die Veränderung der einzelnen Merkmalswerte über die Zeit.
Ausgehend von den aggregierten framebasierten Merkmalen sowie den globalen Merkmalen wird im Folgenden der Merkmalsvektor V_s für einen Klang s generiert. Standardmäßig werden alle Merkmale verwendet. Es ist ebenso möglich, nur eine Auswahl der Merkmale in den Merkmalsvektor aufzunehmen. Die 13 MFCC Koeffizienten werden jeweils als ein Merkmal gerechnet. Wurden alle Merkmale extrahiert, besteht der Merkmalsvektor aus 69 Audiomerkmalen: Fünf globale Merkmale und 16 framebasierte Merkmale, die jeweils anhand von vier Kennwerten aggregiert werden.

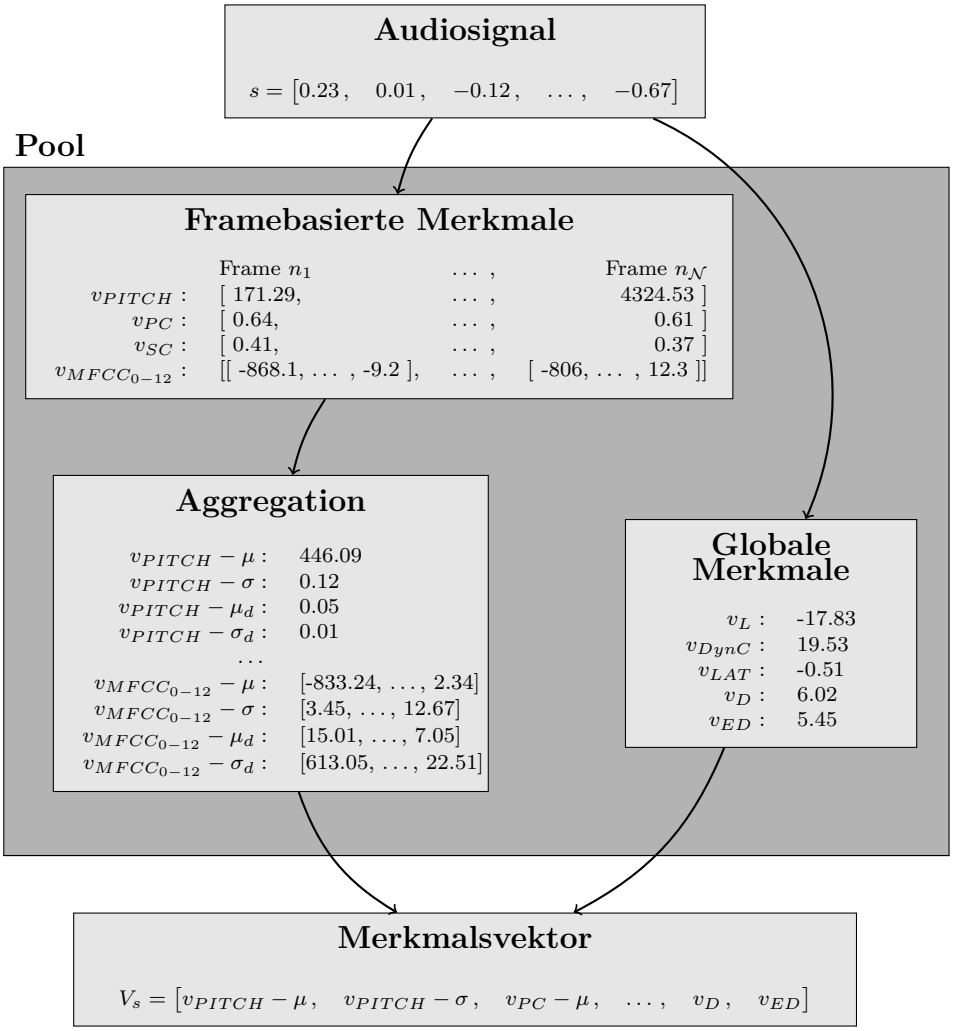


Abbildung 3.7: Schematische Darstellung der Generierung des Merkmalsvektors aus globalen und aggregierten framebasierten Audiomerkmalen. Die Zahlen gehören zu einem fiktiven Beispiel, sind aber realistisch.

3.4 Suche

Die Merkmalsvektoren V_s liegen innerhalb eines $\mathcal{V}$-dimensionalen Vektorraums $\mathbb{R}^{\mathcal{V}}$, wobei $\mathcal{V}$ die Anzahl der Merkmale darstellt. Jeder Merkmalsvektor V_s repräsentiert einen Klang s innerhalb dieses Vektorraums.(Vector Space Model) Der bei der Suchanfrage spezifizierte Beispielklang q wird durch einen entsprechenden Merkmalsvektor V_q repräsentiert. Gegeben V_q sollen nun die k Merkmalsvektoren V_s innerhalb des Vektorraums gefunden werden, die V_q am ähnlichsten sind. Die Distanz zwischen den Vektoren wird als Maß für ihre Ähnlichkeit angenommen. Für die Realisierung der Suche wurde in der vorliegenden Arbeit die Python Bibliothek scikit-learn [82] verwendet. Das Ergebnis der Suche ist eine Liste von k Klängen, die aufsteigend nach Distanz zum Beispielklang sortiert sind.

3.4.1 Distanzmaß

Zur Bestimmung der Ähnlichkeit zweier Vektoren wurde in der vorliegenden Arbeit die euklidische Distanz verwendet (class DistanceMetric). Die euklidische Distanz [59, S. 114] zwischen Vektor V_q eines Beispielklangs und eines beliebigen Vektors V_s ist definiert als

$$d_E = \sqrt{\sum_{v=v_1}^{v_{\mathcal{V}}} \left(V_q(v) - V_s(v)\right)^2}, \tag{3.14}$$

wobei v ein Merkmal und $\mathcal{V}$ die Anzahl aller Merkmale im Vektor sind. Die Anzahl an Rechenoperationen pro Distanzberechnung hängt folglich von $\mathcal{V}$ ab.
Wie in Abschnitt 3.2 deutlich wurde, liegen die Merkmale in verschiedenen Wertebereichen (siehe auch Abbildung 3.7). Alle Merkmale sollen unabhängig von ihrem Wertebereich und Varianzen den gleichen Einfluss auf die Berechnung der Distanz haben. Um dies zu gewährleisten, werden die Merkmalsvektoren vor dem Aufbau des Suchraums standardisiert. Die Standardisierung erfolgt anhand der Funktion scale[19] aus dem Modul preprocessing von scikit-learn. Nach der Standardisierung besitzen alle Merkmale den Mittelwert 0 sowie einheitliche Varianz. Ohne Standardisierung würden Merkmale mit verhältnismäßig hoher Varianz einen größeren Einfluss auf die Distanz ausüben.

[19] Vgl. Onlinedokumentation (`http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.scale.html`, 28. Mai 2016).

3.4.2 k-Nearest-Neighbors-Algorithmus

Für die Berechnung des Suchraums muss zunächst eine $\mathcal{S} \times \mathcal{V}$ Matrix aus allen Merkmalsvektoren erstellt werden. Als Suchalgorithmus wurde in der vorliegenden Arbeit die Implementation des *k-Nearest-Neighbors-Algorithmus* (KNN) von scikit-learn verwendet.[20] Das Verfahren von KNN besteht lediglich darin, eine gegebene Anzahl von nächstliegendsten Punkten zu einem gegebenem Punkt zu finden. Im KNN von scikit-learn sind drei verschiedene Algorithmen zur Realisierung implementiert. Der naive Ansatz zur Lösung des Problems ist, zunächst alle paarweise Distanzen zwischen gespeicherten Merkmalsvektoren zu berechnen. Bei $\mathcal{S}$ Klängen und $\mathcal{V}$ Merkmalen liegt die Komplexität daher im Bereich $\mathcal{O}(\mathcal{V}\mathcal{S}^2)$. Bei einer Suchanfrage muss die Distanz zwischen dem Beispielklang und jedem gespeicherten Klang berechnet werden, wofür jeweils $\mathcal{V}$ Rechenoperationen nötig sind (siehe Unterabschnitt 3.4.1). Die Laufzeit einer Suchanfrage liegt folglich bei $\mathcal{O}(\mathcal{V}\mathcal{S})$.
Um die Komplexität zu reduzieren, können die Daten auf effizienteren Datenstrukturen gespeichert werden. Scikit-learn bietet neben dem naiven Ansatz zwei weitere Varianten von KNN: K-D Tree und Ball Tree. Beide sind Erweiterungen des sogenannten binären Suchbaums. In einem binärem Suchbaum werden Daten so gespeichert, dass ausgehend von einem beliebigem Knoten in der einen Hälfte des Unterbaumes kleinere Werte als der des Knotens gespeichert sind und entsprechend in der anderen Hälfte größere Werte. In K-D Tree und Ball Tree sind in den Knoten Distanzinformation so kodiert, dass daraus ersichtlich wird, in welchem Teilbaum weiter gesucht werden muss. Die Anzahl der zu berechnenden paarweisen Distanzen wird somit reduziert. Die Laufzeit pro Suchanfrage liegt für beide Algorithmen im Bereich $\mathcal{O}(\mathcal{V} \log \mathcal{S})$.[21]
Welche der drei Algorithmen für ein gegebenes Datenset am besten geeignet ist, hängt neben der Größe $\mathcal{S}$ und Dimensionalität $\mathcal{V}$ auch von weiteren Faktoren wie beispielsweise der Anzahl k der gesuchten nächsten Nachbarn ab.[22] Für die vorliegende Arbeit wurde der standardmäßige Parameter 'auto' für die Auswahl des Algorithmus beibehalten. In diesem Fall wird der Algorithmus automatisch in Abhängigkeit von k, $\mathcal{S}$ sowie des verwendeten Distanzmaßes gewählt. Im Kontext der vorliegenden Arbeit heißt das, dass K-D Tree gewählt wird, solange $k < \mathcal{S}/2$.

[20] Vgl. Onlinedokumentation (`http://scikit-learn.org/stable/modules/neighbors.html`, 28. Mai 2016).
[21] Für K-D Tree gilt dies allerdings nur für ungefähr $\mathcal{S} < 20$, vgl. Onlinedokumentation (`http://scikit-learn.org/stable/modules/neighbors.html`, 28. Mai 2016).
[22] Vgl. Onlinedokumentation (`http://scikit-learn.org/stable/modules/neighbors.html`, 28. Mai 2016).

4 Evaluation

Im Kontext der vorliegenden Arbeit ist die zentrale Frage bei der Evaluation folgende: Sind die vom Algorithmus zurückgelieferten Klänge relevant, das heißt dem Beispiel ähnlich? Und damit zusammenhängend: Tauchen relevante Klänge an entsprechend früher Stelle der Ergebnisliste auf? Um diese beantworten zu können, braucht es in erster Linie ein geeignetes Datenset. Darüber hinaus muss innerhalb dieses Datensets zumindest teilweise bekannt sein, welche Klänge als ähnlich gelten können und welche nicht. Da Ähnlichkeit subjektiv und kontextabhängig ist, ist die Relevanz der Suchergebnisse im Grunde nur mithilfe von empirischen Daten zu beurteilen. Diese zu erheben ist wiederum zeitaufwendig, weshalb nur wenige Autoren ihre Evaluation damit stützen. Manche Autoren beschränken sich auf informelle und subjektive Evaluierung [beispielsweise 119, S. 31–32, oder 101, S. 30]. Bei vielen Studien wird vereinfacht davon ausgegangen, dass Klänge ähnlich sind, die in die selbe Klangquellenkategorie fallen. [beispielsweise diverse Studien der Autoren aus 44]. In [54] werden ausschließlich homogene Klänge verwendet, die unterteilt werden und nur diese Teile gelten dann als ähnlich. Nach Kenntnisstand der Autorin gibt es nur drei Studien, in dem Versuchspersonen explizit aufgefordert wurden, die Ähnlichkeit der Klänge eines verwendeten Datensets zu beurteilen [15, 118, 120]. Ausgehend von solchen Daten kann ein sogenannter *Benchmark* (engl. für Maßstab) für die Evaluation des Algorithmus erstellt werden. Aus den genannten Gründen wurde für die Evaluation als Teil der vorliegenden Arbeit ein Datenset sowie ein dazugehöriger Benchmark erstellt. Eine genauere Beschreibung folgt in den nächsten beiden Abschnitten, bevor anschließend auf die eigentliche Evaluation des Algorithmus eingegangen wird. Alle mit der Evaluation zusammenhängenden Daten sind online frei zugänglich.[1]

4.1 Datensätze

Obwohl das Forschungsinteresse bezüglich Umwelt- oder alltäglicher Klänge in den letzten Jahren deutlich zugenommen hat, sind diese nach wie vor im Vergleich zu

[1] Siehe `https://github.com/ESchae/SimilarSoundSearch/tree/master/Evaluation` (6. Juni 2016).

Sprache und Musik ein Randthema innerhalb des Audio Information Retrievals. Dementsprechend haben sich noch wenige Standards etabliert. Inzwischen gibt es zwar ein paar öffentlich zugängliche Datensets[2] und weitere werden in größeren Projekten entwickelt [41, 22]. Allerdings sind die momentan zugänglichen Datensets schwerpunktmäßig auf die automatische Klassifikation anhand der Klangquelle oder Identifikation von Klängen ausgelegt. Diese sind zur Evaluation der vorliegenden Arbeit weniger geeignet. Die Unterschiede zwischen einzelnen Kategorien sind in der Regel deutlich größer als Unterschiede der einzelnen Klänge innerhalb einer Kategorie. Daher wäre vor allem eine kategorische Trennung zu erwarten, was im Kontext der vorliegenden Arbeit nicht beabsichtigt und wenig aussagekräftig wäre [vergleiche 106]. Nach Kenntnisstand der Autorin gibt es kein öffentlich zugängliches Datenset, das für die inhaltsbezogene Suche nach gleichartigen Klängen ausgelegt ist. Ein Datenset zur Evaluation des vorliegenden Algorithmus müsste Klänge beinhalten, die sich unabhängig von der Klangquelle auf unterschiedliche Weise ähneln [vergleiche similar acoustic sounds vs. similar source sounds in 1, S. 65]. Diese zu finden ist jedoch aufwendig. Ein ideales Datenset muss einerseits klein genug sein, um einen Benchmark innerhalb des gegebenen Rahmens der Arbeit aufstellen und den Überblick behalten zu können. Andererseits muss es groß genug sein, um ein Suchszenario unter tatsächlichen Gegebenheiten repräsentieren zu können. Als Kompromiss wurden in der vorliegenden Arbeit daher zwei Datensätze verwendet, die im folgenden beschrieben werden.

4.1.1 Datenset 1

Der erstellte Benchmark bezieht sich auf Datenset 1 (D1).[3] Daher bildet D1 den Schwerpunkt der Evaluation. D1 wurde mithilfe der API von Freesound erstellt. Zunächst wurden zehn Beispielklänge[4] ausgewählt. Ausgehend von diesen wurden jeweils 14 ähnliche Klänge gesucht. Um die Suche nach ähnlichen Klängen zu erleichtern, wurde die Suchfunktion für ähnliche Klänge von Freesound verwendet. Dadurch ist D1 natürlich stark beeinflusst von Freesounds Suchalgorithmus. Diese Voreingenommenheit wird aber durch die Erstellung des Benchmarks relativiert. D1 besteht folglich aus 10 Clustern von 15 Klängen: Jeweils ein Ausgangsklang und 14 dazugehörige Klänge. Die Cluster bilden grob Gruppen ähnlicher Klänge. Es gibt allerdings Ähnlichkeiten zwischen Klängen verschiedener Cluster sowie

[2] Ein guter Überblick findet sich unter `http://www.cs.tut.fi/~heittolt/datasets` (28. Mai 2016).

[3] Auf das Datenset sowie mit dem Datenset zusammenhängende Daten kann frei unter `https://github.com/ESchae/SimilarSoundSearch/tree/master/Evaluation/D1` (6. Juni 2016) zugegriffen werden.

[4] Dies sind die Beispielklänge aus Abschnitt 3.2, allerdings sind 2, 7, 9 und 10 dort gekürzte Versionen.

Unähnlichkeiten innerhalb einzelner Cluster. Die Auswahl der Klänge sollte möglichst verschiedene Aspekte abdecken, die die Ähnlichkeit von Klängen beeinflussen können. Eine tabellarische Übersicht über die 150 Klänge aus D1 findet sich im Anhang in Tabelle A.2.

4.1.2 Datenset 2

D1 ist mit 150 Klängen verhältnismäßig klein. Für die Evaluation des Algorithmus wurde aus zwei Gründen zusätzlich ein größeres Datenset verwendet: Einerseits waren die Klänge in D1 möglicherweise zu unausgewogen oder die Auswahl zu sehr verfälscht durch die Suche von Freesound. Aus rein technischer Perspektive muss andererseits die Effizienz des Algorithmus auf einer größeren Menge von Daten geprüft werden. Das Datenset 2 (D2) ist ein Datenset aus 250.000 Klängen. D2 ist unter dem Namen ESC-US öffentlich zugänglich [87]. D2 ist als Test-Datenset für unüberwachte Lernalgorithmen gedacht. Alle Klänge sind 5 s lang, stammen von Freesound und sind dort als „field recording“ getagged. D2 wird hauptsächlich zur Evaluation der Skalierbarkeit des Algorithmus auf größere Datenmengen verwendet. Das Erstellen eines Benchmarks für ein Datenset solcher Größe war im Rahmen dieser Arbeit nicht möglich. Anstelle einer standardisierte Evaluation wie in Abschnitt 4.3 werden daher lediglich die Suchergebnisse von mehreren zufällig gewählten Beispieldateien von der Autorin begutachtet. Die interessantesten Beobachtungen werden in Abschnitt 4.4 zusammengefasst.

4.2 Benchmarkerstellung über Crowdsourcing

In der Audio Information Retrieval Forschung herrscht grundsätzlich ein Mangel an zugänglichen Benchmarks [73, S. 81–82]. Dies gilt insbesondere für den Bereich der Umwelt- und alltäglichen Klänge. Nach Kenntnisstand der Autorin gibt es keinen öffentlich zugänglichen Benchmark, der für die vorliegende Arbeit geeignet gewesen wäre. Daher wurde ausgehend von den Klängen aus D1 ein Benchmark mittels Crowdsourcing erstellt.[5] Crowdsourcing ist eine komfortable Möglichkeit Daten zu erheben. Aufgaben werden über das Internet gestellt und von weltweit registrierten Nutzern gegen Bezahlung bearbeitet. Crowdsourcing wird seit ein paar Jahren zunehmend zur Evaluation von Information Retrieval Systemen eingsetzt [55]. Für die vorliegende Arbeit wurde die Crowdsourcing Plattform Crowdflower[6] verwendet.

[5] Alle mit der Erstellung des Benchmarks zusammenhängenden Daten sind frei unter `https://github.com/ESchae/SimilarSoundSearch/tree/master/Evaluation` zugänglich.

[6] Siehe `https://www.crowdflower.com/` (5. Juni 2016).

4.2.1 Ablauf

Im folgenden wird der Ablauf der Datenerhebung auf Crowdflower genauer beschrieben. Ergebnisse der Rahmenbedingungen des Ablaufs wie beispielsweise Angaben zu den Teilnehmern, Dauer oder Kosten werden direkt in den entsprechenden Abschnitten berichtet. Die Beschreibung der eigentlichen Studienergebnisse erfolgt anschließend in Unterabschnitt 4.2.2.

Daten Die zehn Beispielklänge aus D1 wurden jeweils mit den restlichen 149 Klängen aus D1 verglichen. Insgesamt bestand das Datenset auf Crowdflower folglich aus 1490 paarweisen Vergleichen. Für jeden der 1490 Vergleiche wurden 5 Beurteilungen gesammelt. Um Ladezeiten gering zu halten, wurde von jedem Klang die Vorschauversion von Freesound in niedrigerer Qualität verwendet. In Abhängigkeit des verwendeten Browsers hörten die Teilnehmer entweder die Version im MP3 oder Ogg Dateiformat.[7] Den Teilnehmern wurde nicht explizit verboten Teile eines Klangs zu überspringen. Sie wurden aber angewiesen, darauf zu achten, keine wichtigen Veränderungen zu überspringen (siehe Abbildung 4.1). Die Annahme dahinter war, dass das typische Nutzerverhalten bei der Suche auch das Überspringen von gleichbleibenden Teilen beinhaltet.

Aufgabe Die Teilnehmer wurden in den Anweisungen über die zugrundeliegende Definition von Ähnlichkeit informiert (s. Anhang oder Link). Die Aufgabenstellung wurde anhand von drei Beispielen verdeutlicht.[8] Die Teilnehmer mussten pro Seite jeweils 5 Aufgaben erfüllen. Eine Aufgabe bestand darin, zwei Klänge anzuhören und zu beurteilen, ob sie ähnlich sind oder nicht (siehe Abbildung 4.1). Die Klangpaare wurden zufällig aus den Crowdflower Daten gewählt. Die Teilnehmer konnten selbst entscheiden, wann sie ihre Teilnahme beenden.

Teilnehmer Die insgesamt $1490 \cdot 5 = 7450$ Beurteilungen wurden von 156 Teilnehmern abgegeben. Ein Teilnehmer tätigte durchschnittlich 61 Beurteilungen. Das Maximum lag bei 362 Beurteilungen pro Teilnehmer, insgesamt tätigten aber nur 7 Teilnehmer mehr als 200 Beurteilungen. Es ist auf Crowdflower nicht möglich, eine verpflichtende Umfrage zur Erhebung von Informationen über die Teilnehmer am Ende anzuhängen. Die Teilnehmer wurden daher gebeten an einer externen Umfrage[9] teilzunehmen, die unter jeder Aufgabe verlinkt wurde (siehe Abbildung 4.1. Im Rahmen von Crowdflower konnte leider nicht garantiert werden, dass jeder

[7] Für die Evaluation des Algorithmus wurden ausschließlich die MP3 Vorschauversionen in niedrigerer Qualität verwendet. Geringe wahrnehmbare Abweichungen der Klangeigenschaften zwischen den Dateiformaten sind nicht auszuschließen, aber an dieser Stelle vernachlässigbar.

[8] Um die Beispiele anzuhören siehe `https://rawgit.com/ESchae/SimilarSoundSearch/master/Evaluation/Crowdflower/Complete/complete_instructions.html` (1. Juni 2016).

[9] Einzusehen unter `https://de.surveymonkey.com/r/YCN36Y5` (1. Juni 2016).

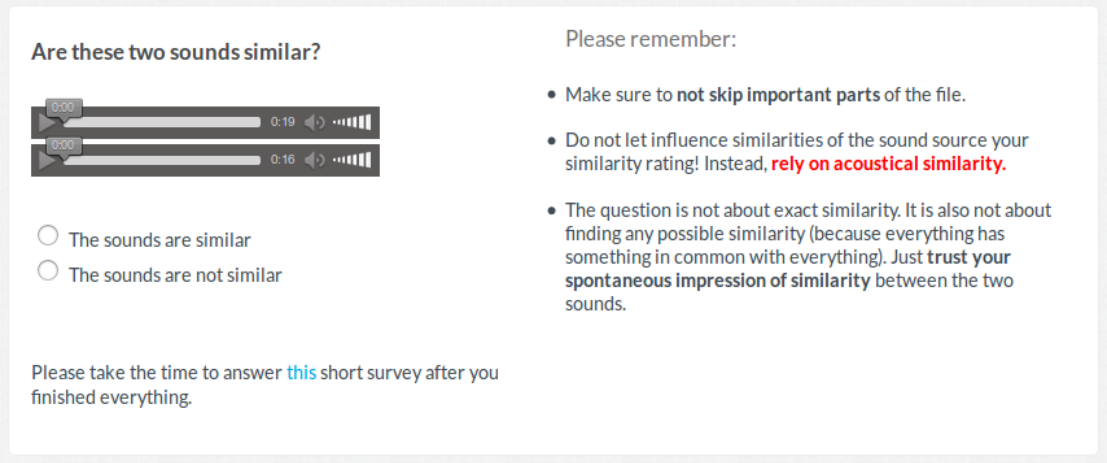


Abbildung 4.1: Aussehen einer Aufgabe auf Crowdflower. Die Ansicht der Klänge wurde bewusst ohne Titel und Wellenform oder ähnlichem gewählt, um jeglichen Kontexteinfluss so gering wie möglich zu halten.

an der Umfrage teilnimmt. Von den 156 Teilnehmern nahmen 58 an der Umfrage teil. Daher sind die Ergebnisse nicht repräsentativ für alle Teilnehmer. Um dafür zu sensibilisieren, welche Aspekte die Beurteilung der Teilnehmer beeinflusst haben könnten, werden die Ergebnisse an dieser Stelle dennoch berichtet. 24 hörten die Klänge über Kopfhörer, 22 über integrierte Lautsprecher und 12 über externe Lautsprecher. Die Klangqualität beurteilten 35 als hoch, 21 als mittel und 2 als niedrig.[10] Bei vier Teilnehmern waren Gehörschädigungen bekannt. 49 gaben an, in der Regel die gesamte Audiodatei angehört zu haben ohne Teile zu überspringen. 6 Teilnehmer gaben an beispielsweise aufgrund des Berufsfeldes häufiger mit Klängen zu arbeiten. Die Mehrheit der Teilnehmer beurteilte die Aufgabe als einfach.[11]

Dauer Die Teilnehmer brauchten durchschnittlich 11 s für die Beantwortung einer Aufgabe.[12] Die gesamte Studie wurde an zwei verschiedenen Tagen durchgeführt. Die Abgabe aller Beurteilungen dauerte insgesamt weniger als 4 h.

Testfragen 77 der 1490 Vergleiche wurden als Testfragen verwendet. Durch sorgfältig gewählte Testfragen kann kontrolliert werden, ob die Teilnehmer die Anweisungen richtig verstanden haben. Darüber hinaus können Betrüger aufgrund von

[10] Natürlich sind diese Angaben subjektiv und dienen nur zur groben Orientierung.
[11] 51 von 58 in der externen Umfrage und innerhalb einer internen Umfrage von Crowdflower, an der sich 45 Teilnehmer beteiligten, lag das Ergebnis bei 4,2 von 5 Punkten unter dem Stichwort „Ease Of Job“.
[12] Der Wert wurde von Crowdflower als Durchschnitt des Interquartils berechnet, das heißt 25% der höchsten und niedrigsten Werte wurden nicht miteingerechnet.

ungewöhnlichem Antwortverhalten gefiltert werden. Als Testfragen wurden 38 relativ unähnliche und 39 relativ ähnliche Klangpaare gewählt.[13]
Jeweils 5 zufällig gewählte Testfragen wurden den Teilnehmern vor Beginn der eigentlichen Studie gestellt. Die 5 ersten Testfragen gehören zum Quiz Mode von Crowdflower. Anhand des Quiz lernen die Teilnehmer den Ablauf kennen und können sich entscheiden, ob sie an der Studie teilnehmen. Ein Teilnehmer muss 3 der 5 Testfragen im Quiz richtig beantworten, um zur Studie zugelassen zu werden. Am Quiz nahmen 233 Teilnehmer teil, davon bestanden 206. 163 der 206 zugelassenen Teilnehmer nahmen anschließend auch an der Studie teil.
In der Studie selbst war auf jeder Seite eine der 5 Fragen eine Testfrage. Die Testfrage war für die Teilnehmer nicht als solche kenntlich. Wenn Teilnehmer eine Testfrage falsch beantworteten, wurden sie darüber informiert. Die richtige Lösung wurde mit Begründung angegeben. Sobald ein Teilnehmer mehr als 70% der Testfragen falsch beantwortet hatte, wurden er und seine bis dahin gefällte Beurteilungen von der Studie ausgeschlossen. Von den 163 Teilnehmern wurden im Verlauf der Studie 7 ausgeschlossen.
Die Teilnehmer hatten die Möglichkeit, das Ergebnis einer Testfrage anzufechten. 52 der 77 Testfragen wurden von Teilnehmern falsch beantwortet. Davon wurden 20 von Teilnehmern angefochten. 13 der Testfragen wurden von 3% oder weniger der Teilnehmer angefochten, die diese Frage falsch beantwortet hatten. 5 der Testfragen wurden von 5–7% der Teilnehmer angefochten. Nur 2 wurden von mehr Teilnehmern angefochten (14% und 23%).[14] Von den 20 angefochtenen Testfragen waren 16 als Beispiele für ähnliche Klänge gedacht. Bei den Anfechtungen zeigt sich folgende Tendenz: Einerseits wurden Testfragen angefochten mit relativ unähnlichen Klängen, die aber die gleiche Klangquelle hatten, beispielsweise zweierlei Miauen von verschiedenen Katzen. Andererseits wurden Testfragen angefochten mit relativ ähnlichen Klängen, die aber eindeutig verschiedene Klangquellen hatten, beispielsweise Uhrticken und Tastaturtippen. An den problematischen Testfragen zeigt sich die Schwierigkeit, Testfragen zu generieren, die kritische Aspekte der Aufgabenstellung kontrollieren und trotzdem nicht zu subjektiv sind.

Kosten Die Teilnehmer erhielten für jede vollständig beantwortete Seite 0,10 $. Teilnehmer, die während der Studie ausgeschlossen wurden, wurden für die bis

[13] Die Beurteilung oblag der Autorin und war zwangsweise subjektiv. Es wurde darauf geachtet möglichst eindeutige Beispiele zu finden, die dennoch aussagekräftig sind. Dies war insbesondere bei den Beispielen für ähnliche Fragen schwierig, dementsprechend wurden diese auch häufiger angefochten (siehe weiter unten).

[14] Letztere wurde aufgrund der vielen Anfechtungen im Verlauf der Studie deaktiviert. Durch die Deaktivierung wurden Beurteilungen von 4 Teilnehmern, die bereits von der Studie ausgeschlossen worden waren, nachträglich in die Daten mit aufgenommen. Dadurch hatten am Ende 20 der Vergleiche 6 anstatt 5 Beurteilungen. Um die Einheitlichkeit zu wahren, wurde bei diesen jeweils die letzte getätigte Antwort wieder heraus genommen.

dahin getätigten Beurteilungen bezahlt. Die Kosten für alle Teilnehmer zusammen mit 20% Geschäftsentgelten beliefen sich auf 228,48 $. Zusammen mit drei Testdurchläufen von jeweils ungefähr 15,00 $ betrugen die Gesamtkosten für das Croudsourcing im Rahmen dieser Arbeit 273,72 $.

4.2.2 Ergebnisse

Obwohl jeder Teilnehmer pro Aufgabe nur zwei Antwortmöglichkeiten hatte, resultiert aus allen 5 Beurteilungen pro Vergleich wieder ein abgestufter Score. Im folgenden ist der Ähnlichkeitsgrad zwischen zwei Klängen definiert als Anzahl von Teilnehmern, die diese als ähnlich beurteilt haben. Als ähnlich gelten Klänge mit einem Ähnlichkeitsgrad ≥ 3, andernfalls gelten sie als unähnlich. Die Verteilung der Ähnlichkeitsgrade ist in Abbildung 4.2 dargestellt. An der Verteilung zeigt sich, dass die Entscheidungen nicht zufällig getroffen worden sind. Bei einer Zufallsverteilung wären die mittleren Ähnlichkeitsgrade verstärkt aufgetreten und Randwerte nur selten.

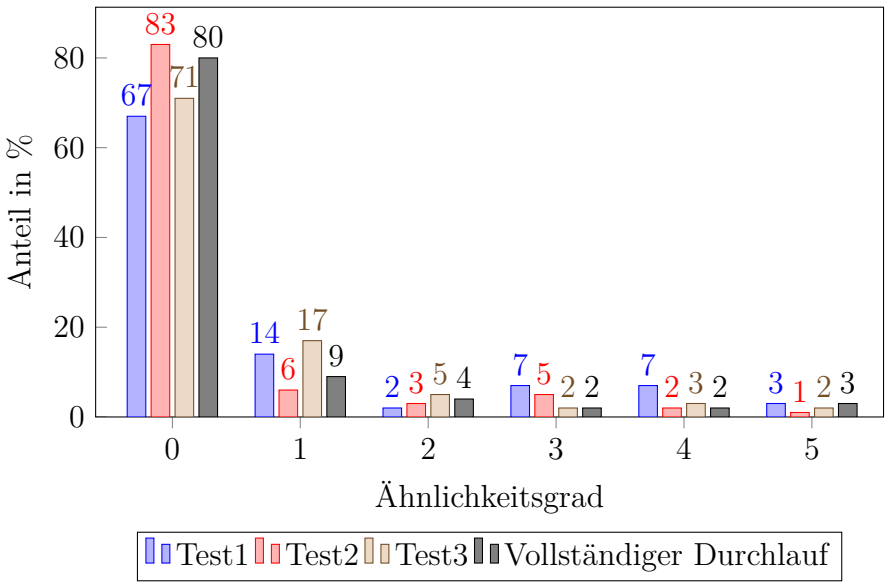


Abbildung 4.2: Verteilung der Ähnlichkeitsgrade. Die Abbildung zeigt den prozentualen Anteil der Ähnlichkeitsgrade innerhalb der drei Testdurchläufe und des vollständigen Durchlaufs. Die Testläufe bestanden jeweils aus 100 der 1490 Vergleiche des vollständigen Durchlaufs und unterschieden sich nur geringfügig in der Aufgabenstellung (siehe erster Punkt Unterabschnitt 4.2.3). Deutlich zu sehen ist, dass hauptsächlich eindeutig unähnliche Klangpaare im Datenset enthalten sind.

Ausgehend von den Clustern aus D1 ist die naive Annahme, dass Klänge innerhalb eines Clusters als ähnlich gelten und Klänge aus verschiedenen Clustern als unähnlich. In diesem Fall wären genau 140 der 1490 Klangpaare ähnlich, nämlich jeweils die Kombinationen zwischen Beispielklang und den 14 Klängen des zugehörigen Clusters. Dies entspräche einem Anteil von 9% ähnlichen Klängen. Die Verteilung aus Abbildung 4.2 deckt sich zwar mit dieser naiven Annahme. Bei einer genauen Betrachtung der Ergebnisse zeigen sich aber deutliche Abweichungen: Es gibt sowohl Unähnlichkeiten innerhalb von Clustern als auch Ähnlichkeiten zwischen Clustern. Die Abweichungen sind in Abbildung 4.3 dargestellt. Es gibt Ähnlichkeiten zwischen Beispielklängen, beispielsweise zwischen Hydrant und Regen. Beides sind Klänge mit starkem Rauschanteil. Konsequenterweise besitzt Hydrant ähnliche Klänge im Cluster von Regen und Regen ähnliche Klänge im Cluster von Hydrant. Die Clusterstruktur ist bei manchen Klängen gut erhalten, beispielsweise bei Regen, Hydrant und Tür, letzterer besteht aus Klängen mit konkreten Tonhöhen, die alle sehr ähnlich sind. In anderen Fällen, beispielsweise bei Spielzeug und Grollen, ist keine Clusterstruktur erkennbar. Das heißt diese Beispielklänge besitzen kaum ähnliche Klänge innerhalb von D1. In jedem Cluster gibt es mindestens einen unähnlichen Klang. In den Clustern Spielzeug, Lachen, Grollen und Schrei wurden mindestens die Hälfte aller Klänge als unähnlich beurteilt. Manche Klänge sind keinem der Beispielklänge ähnlich. Dies sind hauptsächlich Aufnahmen von Sprache (9. und 12. Klang aus Katze, 5. Klang aus Spielzeug), musikalische Patterns (11. Klang aus Spielzeug, 9. Klang aus Feuerwerk) oder lange Aufnahmen von Alltagsszenen (8. Klang aus Katze, 7. Klang aus Grollen). Insgesamt sind die Beurteilungen gut nachvollziehbar mit wenigen Ausnahmen, beispielsweise dem 12. Klang aus Hydrant, der an sich auch einen relativ hohen Rauschanteil besitzt.
Mit den Ergebnissen wird automatisch der Algorithmus von Freesound evaluiert. Es zeigt sich, dass mithilfe der Suchfunktion von Freesound in einigen Fällen nur wenige ähnliche Klänge gefunden wurden. Es kann nicht eindeutig beurteilt werden, ob dies an der Suchfunktion von Freesound liegt oder ob möglicherweise im Klangarchiv von Freesound keine ähnlicheren Klänge enthalten sind. Allerdings stehen relevante Ergebnisse an sehr unterschiedlichen Positionen innerhalb der Ergebnislisten. Es ist daher sehr wahrscheinlich, das weitere relevante Ergebnisse an späteren Positionen (> 14) stehen und folglich nicht gefunden wurden.

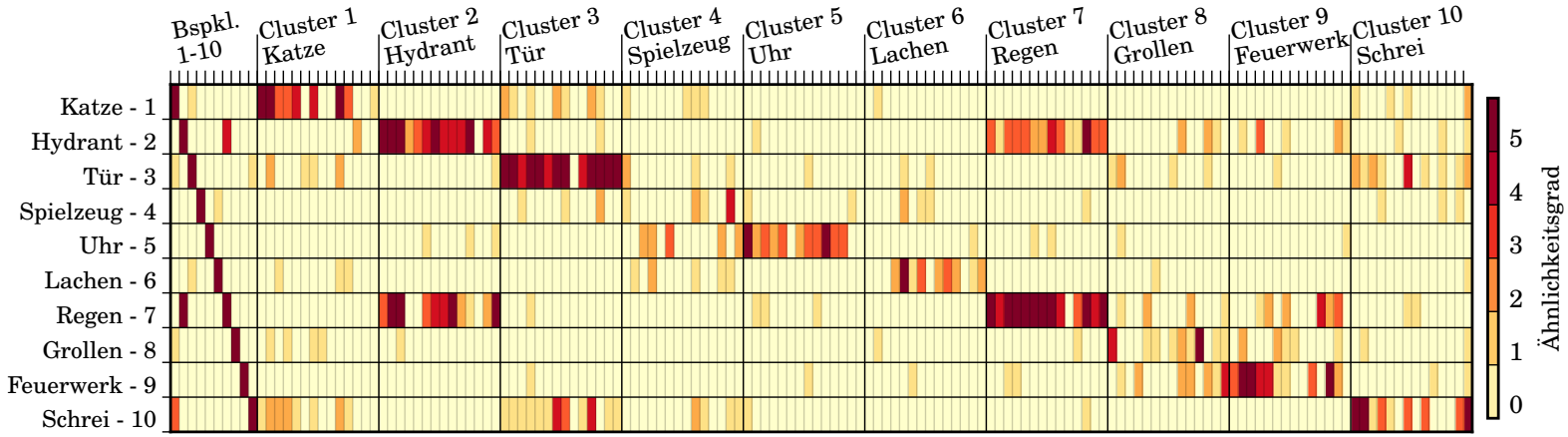


Abbildung 4.3: Darstellung der Crowdsourcing Ergebnisse. Jede Spalte repräsentiert einen der 150 Klänge aus D1. Die Reihenfolge der Spalten entspricht von links nach rechts der Reihenfolge der Klänge in der Tabelle A.2 aus dem Anhang: In den ersten zehn Spalten stehen die Beispielklänge (Bspkl.) 1–10 und anschließend die zehn zugehörigen Cluster mit jeweils 14 Klängen. Die 14 Klänge innerhalb eines Cluster sind aufsteigend entsprechend der Distanz von Freesound sortiert. Das heißt, der Klang in der Spalte ganz links innerhalb eines Clusters ist laut Freesound dem zugehörigen Beispielklang am ähnlichsten. Der Klang in der Spalte ganz rechts innerhalb eines Clutsers ist dem zugehörigen Beispielklang laut Freesound am wenigsten ähnlich. In jeder Reihe sind die Ähnlichkeitsgrade zwischen einem Beispielklang und allen 150 Klängen aus D1 farblich dargestellt. Der Ähnlichkeitsgrad zwischen einem Beispielklang aus Reihe i und einem Klang aus Spalte j ist anhand des Farbwerts in Zelle x_{ij} dargestellt. Die dunkelrote Diagonale in den ersten zehn Spalten zeigt beispielsweise, dass jeder Beispielklang zu sich selbst einen Ähnlichkeitsgrad von 5 besitzt. In D1 sind hauptsächlich unähnliche Klänge enthalten, erkennbar daran, dass die meisten Zellen den Ähnlichkeitsgrad 0 besitzen.

4.2.3 Kritische Anmerkungen

Die Durchgeführte Studie auf Crowdflower unterliegt einigen Einschränkungen, die in zukünftigen Untersuchungen verbessert werden könnten. Die wichtigsten sind folgende:

Qualitatives Feedback Ursprünglich war die Idee, neben den bloßen Ähnlichkeitsbeurteilungen auch Informationen über die Gründe der Beurteilung zu erhalten. Wenn bekannt wäre, welche Klangaspekte die Entscheidung besonders beeinflussen, könnte solches Wissen auch zur Verbesserung eines Algorithmus verwendet werden. Allerdings zeigte sich in den Testläufen, dass Crouwdsourcing dafür weniger geeignet ist. Freitextantworten lieferten einerseits kaum aussagekräftige Antworten und sind schwer auszuwerten. Die Alternative wäre, mögliche Antworten per Multiple-Choice-Verfahren anzugeben. Die Wahl geeigneter Klangaspekte für Antwortmöglichkeiten, die die Ähnlichkeitsbeurteilung beeinflusst haben könnten, wäre allerdings selbst wieder ein eigenes Forschungsthema. Für solche Art von Ergebnissen wäre sicherlich eine Studie mit Experten wie beispielsweise Sound Designern sinnvoller gewesen. Abgesehen davon ist nicht auszuschließen, dass Experten die Klangähnlichkeiten anders beurteilt hätten [18]. Je nach Zielgruppe des Suchalgorithmus wäre es unter Umständen von vorn herein aufschlussreicher gewesen, die Studie mit Experten durchzuführen.

Testfragen Möglicherweise waren die Testfragen zu voreingenommen. Da es keine Begründung des Antwortverhaltens über Freitext oder ähnlichem gab, waren die Testfragen allerdings die einzige Möglichkeit zu überprüfen, ob die zugrundeliegende Definition von Ähnlichkeit verstanden und befolgt wurde. Die Testfragen müssten in Zukunft verbessert werden, um noch weniger Anfechtungen zu erhalten. Unter Umständen wären im Zuge dessen auch bessere Anweisungen nötig. Möglicherweise wäre auch ein Verhältnis von 9 Aufgaben plus 1 Testfrage pro Seite ausreichend zur Qualitätskontrolle.

Häufung der Beispielklänge Im Gesamten war die Aufgabe trotz Randomisierung der Reihenfolge unausgewogen, da nicht alle $150 \cdot 150 = 22500$ Vergleiche zwischen den Klängen durchgeführt wurden. Dadurch traten die zehn Beispielklänge im Vergleich zu den restlichen Klängen gehäuft auf, was möglicherweise Beurteilungen beeinflusst haben könnte.

4.3 Evaluation mittels D1

Mithilfe der Crowdsourcing Ergebnisse wurde ein Benchmark erstellt. Dieser ist zusammen mit den anderen Daten der Evaluation online frei zugänglich.[15] Im Benchmark wird jedem Beispielklang eine Liste mit den restlichen 149 Klängen aus D1 zugeordnet. Zusammen mit jedem Klang steht dort der zugehörige Ähnlichkeitsgrad (siehe Unterabschnitt 4.2.2) zum Beispielklang. Die Liste ist absteigend nach Ähnlichkeitsgrad sortiert. Je früher ein Klang in der Liste auftaucht, als desto relevanter wird er interpretiert. Der Benchmark repräsentiert sozusagen die ideale Ergebnisliste, die ein Suchalgorithmus für jeden der zehn Beispielklänge liefern sollte.

4.3.1 Verglichene Algorithmen

Für die Evaluation wurden die folgenden vier Algorithmen verglichen:

Alle Merkmale Der vorliegende Algorithmus wie in Kapitel 3 beschrieben, das heißt, mit allen in Abschnitt 3.2 beschriebenen Audiomerkmalen.

Nur MFCC Der selbe Algorithmus, allerdings nur mit den ersten 13 MFCC als Audiomerkmale. Die MFCC werden in der Forschung häufig als naive Baseline verwendet, da sie trotz mangelndem Wahrnehmungsbezug relativ gute Ergebnisse liefern.

Ohne MFCC Der vorliegende Algorithmus wie in Kapitel 3 beschrieben, allerdings ohne MFCC. Diese Variante wurde verwendet, um den Einfluss der MFCC auf die Ergebnisse des vorliegenden Algorithmus zu überprüfen.

Freesound Der Algorithmus von Freesound (siehe Unterabschnitt 3.1.2). Mithilfe der kombinierten Suchfunktion der Freesound API[16] konnten die Distanzen zwischen allen 1490 Klangpaaren erhalten werden.

Jedem der vier Algorithmus wurden alle zehn Beispielklänge als Suchanfragen gestellt. Alle Algorithmen liefern für jeden der Beispielklänge eine Liste von Klängen zurück, die aufsteigend nach zunehmender Distanz zum Beispielklang sortiert ist. Für jeden Algorithmus wurden die zehn Ergebnislisten der Beispielklänge mit der jeweiligen Ergebnisliste des Benchmarks verglichen.

[15] Siehe `https://github.com/ESchae/SimilarSoundSearch/tree/master/Evaluation` (5. Juni 2016), der Benchmark befindet sich im Unterordner Evaluation_with_D1.

[16] Um die Distanz zwischen zwei beliebigen Klängen zu erhalten, muss der erste Klang als Suchanfrage und der zu vergleichende Klang als Filter verwendet werden. Vgl. Onlinedokumentation (`http://www.freesound.org/docs/api/resources_apiv2.html`, 3. Juni 2016).

4.3.2 Abbildung der Distanzen auf Ähnlichkeitsgrade

Die vier Algorithmen liefern für jedes Klangpaar einen Distanzwert zurück. Die Distanzen liegen in unterschiedlichen kontinuierlichen Wertebereichen. Für die Evaluation müssen die Distanzwerte der Algorithmen auf einen ganzzahligen Ähnlichkeitsgrad zwischen 0 und 5 abgebildet werden. Die Verteilungen der Distanzen für alle zehn Ergebnislisten pro Algorithmus sind in Abbildung 4.4 dargestellt. Oben links in der Abbildung sind zum Vergleich die Ähnlichkeitsgrade des Benchmarks abgebildet. Deutlich erkennbar ist, dass die Ähnlichkeitsgrade im Benchmark sehr schnell abfallen. Ähnnlichkeitsgrade > 0 finden sich nur im ersten Drittel der Ergebnislisten. Die Verteilung der Distanzen der vier Algorithmen ist abgesehen von unterschiedlichen Wertebereichen relativ ähnlich. Am Anfang und Ende der Ergebnislisten nehmen die Distanzen stärker zu. In der Mitte ist die Steigung sehr gering und weitestgehend linear. Auffällig bei Freesound ist ein Knick nach den ersten 14 Ergebnissen. Dieser lässt sie wie folgt erklären: Für jeden Beispielklang stammen die ersten 14 Klänge aus der eigentlichen Ergebnisliste der inhaltsbezogenen Suche auf Freesound. Die nachfolgenden Klänge aus D1 entsprechen nicht den Klängen, die auf Freesound nachfolgend in der Ergebnisliste stehen würden. Zum Vergleich ist in Abbildung 4.4 unten rechts die Verteilung der Distanzen der eigentlichen Ergebnislisten der Beispielklänge auf Freesound dargestellt. Die Ergebnisse 15–149 sind in diesem Fall Klänge, die nicht in D1 enthalten sind.

Für die Skalierung wurden folgende zwei Varianten verwendet, analog zu [4, S. 6]. In Abbildung 4.5 ist die Verteilung der Ähnlichkeitsgrade nach beiden Arten der Skalierung dargestellt.

Linear Als Maximum wird die maximale Distanz innerhalb einer Ergebnisliste angenommen. Jede Distanz wird zunächst durch das Maximum dividiert und anschließend mit 5 multipliziert. Da die Distanzen zunehmen, der Ähnlichkeitsgrad jedoch abnimmt wird anschließend das Ergebnis von 5 subtrahiert. Abschließend wird auf die nächste ganze Zahl gerundet. Nach der linearen Skalierung entspricht die Verteilung der Ähnlichkeitsgrade der Verteilung der Distanzen. Das heißt konkret, die Extremwerte 0 und 5 tauchen selten auf, mittlere Werte dagegen häufig.

Logarithmisch Jede Distanz wird ebenfalls durch das Maximum dividiert. Anschließend wird das Ergebnis mit 2^5 multipliziert, davon der Logarithmus zur Basis 2 genommen und das Ergebnis von 5 subtrahiert. Um zu reflektieren, dass im Benchmark hauptsächlich Ähnlichkeitsgrad 0 vorhanden ist, wurden Ergebnisse ≥ 1 zur nächsten ganzen Zahl gerundet und Ergebnisse < 1 auf 0 abgerundet. Die logarithmische Skalierung hat eine Verteilung der Ähnlichkeitsgrade zur Folge, die eher der Verteilung der Ähnlichkeitsgrade des Benchmarks entspricht.

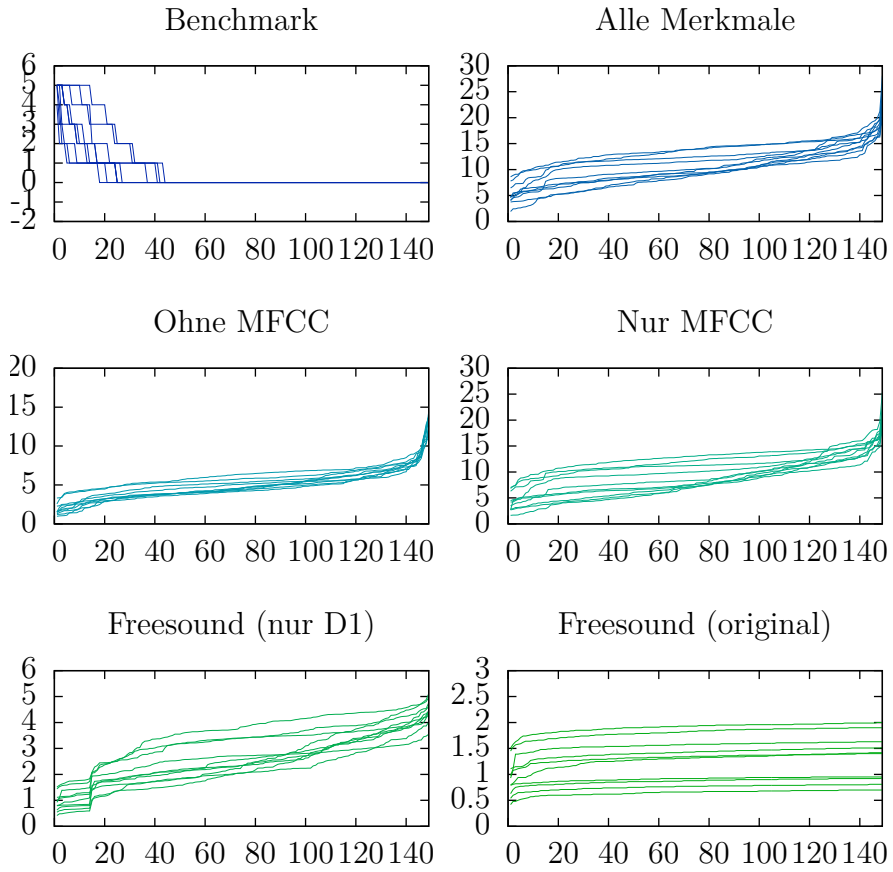


Abbildung 4.4: Verteilung der Distanzwerte. Für jeden Algorithmus sind die Distanzen der Ergebnislisten aller zehn Beispielklänge dargestellt. Die Linien zeigen die Distanzen zwischen einem Beispielklang und allen 149 restlichen Klängen aus D1. Auf der x-Achse sind die 149 Ergebnisse abgetragen, sortiert nach zunehmender Distanz zum Beispielklang. Auf der y-Achse sind die jeweiligen Distanzen zwischen Ergebnis und Beispielklang dargestellt. Die Ähnlichkeitsgrade im Benchmark nehmen mit zunehmender Distanz zum Beispielklang ab, die Distanzwerte jedoch zu. Unten links sind die Freesound Distanzen der Beispielklänge zu den Klängen aus D1 dargestellt. Unten rechts sind die Freesound Distanzen dargestellt, die man bei einer originalen Suchanfrage der Beispielklänge auf Freesound für die ersten 150 Ergebnisse erhalten würde.

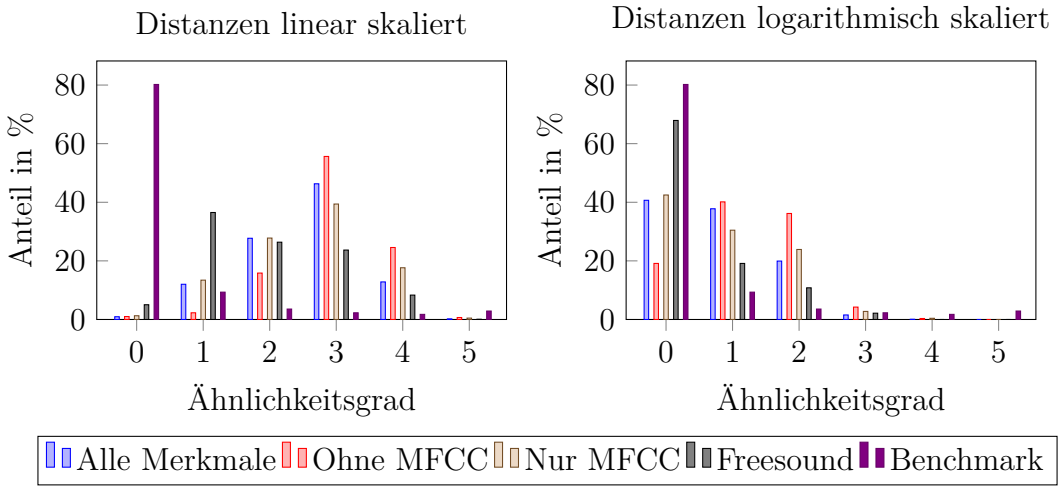


Abbildung 4.5: Verteilung der Ähnlichkeitsgrade nach linearer (links) und logarithmischer (rechts) Skalierung der Distanzwerte im Vergleich zur Verteilung der Ähnlichkeitsgrade im Benchmark.

4.3.3 Evaluationsmaße

Die Ergebnislisten wurden auf zwei verschiedene Arten evaluiert: Einerseits *score-based*, das heißt die über Skalierung erhaltenen Ähnlichkeitsgrade werden mit den Ähnlichkeitsgraden des Benchmarks verglichen. Andererseits *rank-based*, das heißt die Reihenfolge der Suchergebnisse wird evaluiert.

Genauigkeit Die Genauigkeit (engl. accuracy) beschreibt den Anteil an Klangpaaren, bei denen der Ähnlichkeitsgrad des Algorithmus höchstens um δ vom Ähnlichkeitsgrad im Benchmark abweicht [4, S. 7].

Durchschnittliche Abweichung Die durchschnittliche Abweichung (engl. average score deviation) ist der Durchschnitt von allen absoluten Abweichungen zwischen Ähnlichkeitsgraden innerhalb einer Ergebnisliste[4, S. 7].

nDCG Unter dem Rang (engl. rank) wird im folgenden die Position eines Ergebnis innerhalb der Ergebnisliste verstanden. Gegeben eine Ergebnisliste, wird jedem Ergebnis der entsprechende Ähnlichkeitsgrad des Benchmarks zugeordnet. Der *Cumulative Gain* (CG) ist definiert als Summe aller dieser Ähnlichkeitsgrade bis zu einem bestimmten Rang r [48, S. 42]. Relevante Ergebnisse sollten möglichst früh in der Ergebnisliste stehen. Beim *Discounted Cumulative Gain* (DCG) wird die Reihenfolge der Ergebnisse berücksichtigt, indem die Ähnlichkeitsgrade a mit

dem Logarithmus der Position i innerhalb der Ergebnisliste gewichtet werden:

$$\mathrm{DCG}_r = a_1 + \sum_{i=2}^{r} \frac{a_i}{\log_2(i)} \,. \tag{4.1}$$

Je später das Ergebnis in der Ergebnisliste auftaucht, das heißt je größer i, desto weniger trägt dessen Ähnlichkeitsgrad zum DCG bei. Der DCG wird beeinflusst durch die Länge der Ergebnisliste sowie die Ähnlichkeitsgrade innerhalb der Ergebnisse. Um Ergebnisse verschiedener Ergebnislisten besser miteinander vergleichbar zu machen, kann der sogenannte *normalized Discounted Cumulative Gain* (nDCG) berechnet werden. Dieser ist definiert als

$$\mathrm{nDCG_r} = \frac{\mathrm{DCG_r}}{\mathrm{IDCG_r}} \,, \tag{4.2}$$

wobei IDCG der sogenannte *ideale* DCG ist. Um den IDCG zu erhalten, wird die Ergebnisliste anhand der zugeordneten Ähnlichkeitsgrade des Benchmarks sortiert und auf dieser der DCG berechnet.

4.3.4 Ergebnisse

Die Ergebnisse der Evaluation sind in Tabelle 4.1 dargestellt. Im Folgenden werden zunächst die Ergebnisse von Genauigkeit und durchschnittlicher Abweichung und anschließend die Ergebnisse von nDCG detaillierter beschrieben. Zur Berechnung der Signifikanz wurde der R-Test verwendet. Als signifikant gelten Unterschiede, bei denen der p-Wert des R-Tests unterhalb des α-Niveaus von 5% liegt. Es werden nur die wichtigsten Ergebnisse berichtet. Eine ausführliche Übersicht über alle paarweise berechneten p-Werte findet sich im Anhang (Abschnitt A.2).

4.3.4.1 Genauigkeit und durchschnittliche Abweichung

Wie zu erwarten war, führt die logarithmische Skalierung für alle Algorithmen zu besseren Ergebnissen. In Abbildung 4.5 wurde deutlich, dass die Verteilung der Ähnlichkeitsgrade von Freesound der des Benchmarks am nächsten kommt. Dementsprechend ist die Genauigkeit von Freesound höher als die der anderen Algorithmen.
Bei logarithmischer Skalierung liegt die Genauigkeit aller Algorithmen für $\delta = 2$ bei 97%. Dieses Ergebnis lässt sich relativ einfach erklären: Nach der logarithmischen Skalierung liegen fast alle Ähnlichkeitsgrade zwischen 0 und 2. Da im Benchmark die meisten Ähnlichkeitsgrade 0 sind, kann es folglich in den wenigsten Fällen zu einer Abweichung um mehr als 2 kommen.

Tabelle 4.1: Ergebnisse der Evaluation. Jede Reihe zeigt die Ergebnisse der Evaluationsmaße für einen Algorithmus. Die einzelnen Werte sind jeweils der Durchschnitt von allen zehn Beispielklängen. Bei Genauigkeit und durchschnittlicher Abweichungen stehen pro Zelle die Ergebnisse für lineare (links) und logarithmische Skalierung (rechts). Ausgehend von den nicht gerundeten Ergebnissen ist jeweils das beste Ergebnis fett gedruckt.

Algorithmus	Genauigkeit (%)			Durchschnittliche Abweichung	nDCG
	$\delta = 0$	$\delta = 1$	$\delta = 2$		
Ohne MFCC	4 | 24	12 | 67	34 | **97**	0,21 | **0,15**	0,78
Nur MFCC	4 | 43	25 | 78	56 | 97	0,19 | 0,15	0,80
Alle Merkmale	4 | 43	23 | 83	55 | 97	0,18 | 0,15	**0,87**
Freesound	**10 | 65**	**53 | 89**	**80** | 97	**0,17** | 0,16	0,81

Bei der Genauigkeit zeigen sich im wesentlichen zwei Tendenzen: Die Ergebnisse von Freesound sind besser als die der anderen Algorithmen und die Ergebnisse von Ohne MFCC sind schlechter als die der anderen Algorithmen. Die Ergebnisse von Freesound unterscheiden sich in den meisten Fällen signifikant von den anderen Algorithmen ($0.20\% \leq p \leq 2.73\%$). Die einzigen Ausnahmen sind bei logarithmischer Skalierung die Unterschiede für $\delta = 2$ sowie der Unterschied zwischen Freesound und Ohne MFCC für $\delta = 1$ ($p = 6.25\%$). Dass die Verteilung der Ähnlichkeitsgrade von Freesound im Vergleich zu den anderen Algorithmen eher der des Benchmarks ähnelt, war bereits in Abbildung 4.5 gut zu erkennen.
Die Ergebnisse von Ohne MFCC unterscheiden sich bei linearer Skalierung für $\delta = 1$ und $\delta = 2$ signifikant von den anderen Algorithmen ($0.20\% \leq p \leq 0.78\%$). Bei logarithmischer Skalierung ist neben den Unterschieden für $\delta = 2$ lediglich der Unterschied für $\delta = 1$ zu Nur MFCC nicht signifikant. Zwischen Nur MFCC und Alle Merkmale ist nur der Unterschied für $\delta = 1$ bei logarithmischer Skalierung signifikant ($p = 3.91\%$).
Bei der durchschnittlichen Abweichung ist nur der Unterschied zwischen Freesound und Ohne MFCC bei linearer Skalierung signifikant ($p = 3.12\%$). Im Anbetracht der Ähnlichkeitsgrade ist ein Unterschied von 0,04 allerdings sehr gering und daher trotz Signifikanz vernachlässigbar.

4.3.4.2 nDCG

Die durchschnittlichen nDCG Werte sind insgesamt relativ hoch für alle Algorithmen. Bei den Ergebnissen des nDCG ist nur der Unterschied zwischen Alle Merkmale und Ohne MFCC signifikant ($p = 3.32\%$). Die einzelnen nDCG Werte pro Beispielklang und Algorithmus sind in Abbildung 4.6 dargestellt.

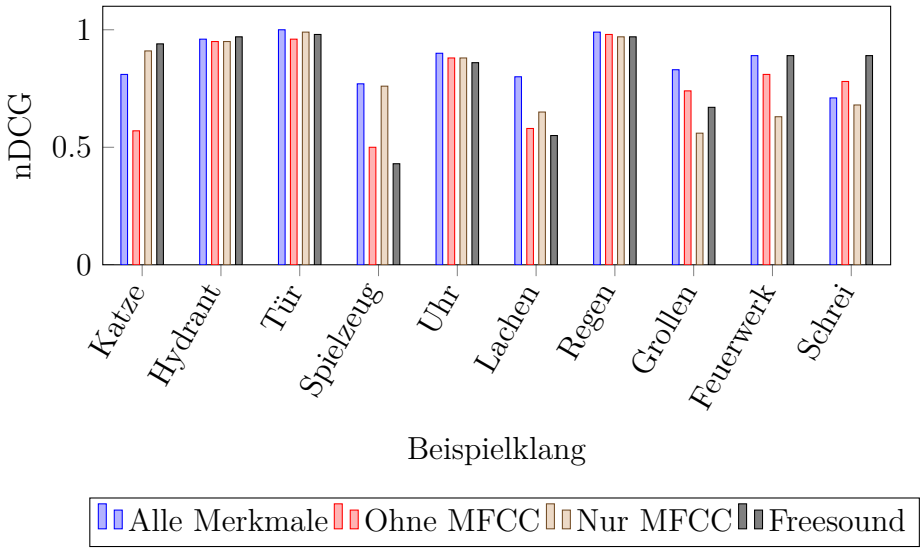


Abbildung 4.6: nDCG pro Beispielklang und Algorithmus.

Bei genauerer Analyse der nDCG Werte pro Beispielklang ergeben sich folgende Tendenzen: Bei vier Beispielklängen (Hydrant, Tür, Uhr und Regen) ist der nDCG aller Algorithmen ungefähr gleich hoch und insgesamt sehr hoch. Hydrant, Tür und Regen sind die Beispielklänge mit den meisten ähnlichen Klängen im Benchmark. In allen drei Ergebnislisten finden sich verhältnismäßig viele Klänge mit einem Ähnlichkeitsgrad von 5 oder 4.[17] Beispielklänge, die im Benchmark relativ wenig ähnliche Klänge besitzen (Spiel, Lachen und Grollen) haben tendenziell einen niedrigeren und uneinheitlicheren nDCG. Laut dem Benchmark gibt es in D1 beispielsweise zu Spielzeug nur einen Klang mit Ähnlichkeitsgrad 5, alle anderen besitzen einen Ähnlichkeitsgrad von ≤ 3 und bei fast allen ist der Ähnlichkeitsgrad 0 (siehe Abbildung 4.3). In diesem Fall bestimmt die Position dieses einen Klangs

[17] Im Cluster von Tür sind hauptsächlich Klänge mit konkreter Tonhöhe enthalten. Die Tonhöhe entspricht in den meisten Fällen der Tonhöhe des Beispielklangs. Hydrant und Regen sind Cluster mit rauschenden Klängen. Beide Aspekte hängen mit der Tonhaltigkeit zusammen und scheinen ein wesentliches Merkmal in der Ähnlichkeitsbeurteilung von Klängen zu sein. Dies zeigte sich unter Anderem in den qualitativen Ergebnissen der Testdurchläufe auf Crowdflower (siehe Unterabschnitt 4.2.3) und deckt sich mit Ergebnissen aus den MDS Studien (siehe Unterabschnitt 2.5.3).

weitestgehend den nDCG. Die Verteilung der Ähnlichkeitsgrade im Benchmark ist ein Einflussfaktor auf die Unterschiede der nDCG Ergebnisse. Sie erklärt aber nicht die unterschiedlichen Ergebnisse der Algorithmen. In den meisten Fällen sind die Ergebnislisten von Alle Merkmale und Nur MFCC sehr ähnlich. In den Ergebnislisten ist für viele Beispielklänge die Clusterstruktur von D1 erkennbar. Da D1 mit Freesounds Suchfunktion erstellt wurde bedeutet das, dass Alle Merkmale und Nur MFCC auch sehr ähnlich zu Freesound sind. Ohne MFCC unterscheidet sich dagegen tendenziell stärker von den anderen Algorithmen.

4.3.5 Diskussion

Insgesamt zeigt sich bei der Analyse der nDCG Werte die selbe Tendenz wie bei Genauigkeit und durchschnittlicher Abweichung. Die drei wesentlichen Ergebnisse der Evaluation werden im folgenden zusammengefasst. Die ersten beiden Ergebnisse decken sich mit den Ergebnissen einer Studie von Aucouturier & Pachet [79, S. 8–9]. Die Autoren haben in dieser Studie einen umfassenden Vergleich verschiedener Algorithmen zur Bestimmung von Klangfarbenähnlichkeit durchgeführt.

Alle Ähnlich Die Ergebnisse der vier Algorithmen sind sehr ähnlich. Vor allem Alle Merkmale und Nur MFCC sind sehr ähnlich. Ohne MFCC schneidet allerdings tendenziell schlechter ab als die anderen drei Algorithmen. Freesound scheint bezüglich Genauigkeit besser zu sein. Die guten Ergebnisse von Freesound bezüglich Genauigkeit sind möglicherweise ein Artefakt des Knicks (siehe Abbildung 4.4) in der Verteilung der Distanzen. Dass die Algorithmen sich in der Evaluation nicht wesentlich voneinander unterscheiden kann auch dadurch beeinflusst worden sein, dass D1 verhältnismäßig klein ist und mit lediglich zehn Beispielklängen evaluiert wurde.

Ausreißer In allen Ergebnislisten sind relativ früh immer wieder unähnliche Klänge enthalten. Diese sind, nach Aucouturier & Pachet „nicht ‚anfechtbar weniger ähnlich', sondern in der Regel *sehr* schlechte Ergebnisse, die objektiv nichts zu tun haben" mit dem Beispielklang [Übersetzung durch Autorin; 79, S. 9]. Wenn man sich die Cluster aus D1 anhört, kann man einen guten Eindruck davon bekommen.

Einfluss der MFCC Nur MFCC und Alle Merkmale führen hauptsächlich zu den gleichen Ergebnissen und Ohne MFCC zu schlechteren Ergebnissen. Daraus lässt sich schlussfolgern, dass die MFCC einen wesentlichen Einfluss auf die Genauigkeit ausüben. Die Leistungsfähigkeit der MFCC deckt sich mit zahlreichen Studien. Der Einfluss der MFCC auf die Ergebnisse von Freesound kann nicht beurteilt werden.[18] Möglicherweise spielt die Auswahl der Audiomerkmale abgesehen

[18] Insgesamt ist der Einfluss einzelner Audiomerkmale bei Freesound durch das Verwenden der PCA nur schwer zu interpretieren.

von den MFCC nur eine untergeordnete Rolle. Es könnte aber genauso sein, dass lediglich die im Rahmen der vorliegenden Arbeit verwendeten Audiomerkmale relevante Wahrnehmungsaspekte nicht ausreichend wiedergeben. Um genauere Aussagen treffen zu können sind weitere Untersuchungen nötig (siehe Abschnitt 4.5).

4.4 Evaluation mittels D2

Mithilfe von D2 wurde einerseits die Skalierbarkeit des implementierten Algorithmus getestet. Andererseits wurde eine qualitative Analyse der Suchergebnisse durchgeführt. Die Ergebnisse von beidem werden im Folgenden beschrieben.

4.4.1 Laufzeit

Die Laufzeit wurde auf einem Rechner mit 8 GB RAM und Intel i5-5200U CPU @ 2.20GHz x 4 Prozessor gemessen. Beim ersten Versuch des Einlesens von D2 wurde ein Speicherleck entdeckt. Dieses stellte sich als internes Leck von Essentia heraus und konnte mithilfe von parallel laufenden Prozessen umgangen werden.[19] Die Komplexität des Algorithmus nimmt proportional zur Anzahl Audiodateien zu. Die Merkmalsextraktion ist der Teil des Algorithmus mit höchster Speicher- und Rechenkomplexität. Die Komplexität der Merkmalsextraktion nimmt proportional mit der Länge einer Audiodatei zu.[20] Beim Einlesen der Audiodateien in die Datenbank werden jeweils die Audiomerkmale von 100 Klängen extrahiert und anschließend in die Datenbank geschrieben. Die Merkmalsextraktion für 100 Audiodateien dauerte durchschnittlich 7 s. Das Einlesen von allen 250.000 Audiodateien aus D2 dauerte ungefähr 5 h. Zum Vergleich: Das Einlesen der 150 Klänge aus D1 dauerte 2 min. Wohlgemerkt beinhaltet D1 mehrere Klänge, die deutlich länger sind als die 5 s langen Audiodateien aus D2. Die verhältnismäßig lange Dauer ist vertretbar, wenn die Audiodateien offline – das heißt vor der eigentlichen Programmausführung – eingelesen werden.
Die Berechnung des KNN-Suchraums dauerte durchschnittlich ungefähr 3 min. Bei D1 dauerte es im Vergleich lediglich 0,1 s. Es ist möglich den einmal aufgebauten Suchraum über die Programmanwendung hinaus zu speichern. Das Laden des gespeicherten Suchraums von D2 benötigt im Schnitt ungefähr 3 s. Ausgehend vom aufgebauten Suchraum dauerte die Beantwortung einer Suchanfrage aus D2

[19] Zum Speicherleck von Essentia siehe `https://github.com/MTG/essentia/issues/405` (4. Juni 2016). Für Details der Umsetzung siehe Quellcode unter `https://github.com/ESchae/SimilarSoundSearch` (4. Juni 2016).

[20] Da für alle Audiodateien die selbe Abtastrate und Framelänge verwendet werden, sind diese Größen vernachlässigbar.

durchschnittlich ungefähr 0,07 s. Wohlgemerkt sind alle Klänge in D2 5 s lang. Die Laufzeit einer Suchanfrage ist auch abhängig von der Länge des verwendeten Beispielklangs, da für diesen zunächst alle Merkmale extrahiert werden müssen. Zum Vergleich: Die Beantwortung der selben Suchanfrage auf D1 dauerte durchschnittlich 0,05 s.

4.4.2 Suchergebnisse

Um die Qualität der Suchergebnisse auf D2 zu überprüfen wurden ungefähr 100 zufällig ausgewählte Audiodateien aus D2 als Suchanfrage gewählt und die Ergebnisse anschließend überprüft. Die 14 ersten Ergebnisse von 26 verschiedenen Suchanfragen sind online anzuhören.[21] Es zeigten sich folgende Tendenzen: Die Klänge der Ergebnislisten sind in den meisten Fällen relativ ähnlich. Zwei Punkte müssen beachtet werden: Erstens sind die Klänge aus D2 alle sogenannte *Field Recordings*, das heißt, es handelt sich um Aufnahmen von Umweltgeräuschen und alltäglichen Szenen. Ein Großteil der Klänge scheint daher relativ ähnlich zu sein. Eine genaue Beurteilung ist natürlich bei 250.000 Klängen nicht möglich. Zweitens handelt es sich hierbei nur um den subjektiven Eindruck der Autorin, der dadurch beeinflusst werden könnte, dass innerhalb einer Ergebnisliste automatisch nach Ähnlichkeiten gesucht wird, selbst wenn diese möglicherweise nicht so offensichtlich in den Klängen enthalten sind. Der Leser sei daher an dieser Stelle ermuntert, sich selbst einen Eindruck der Ergebnisse zu verschaffen.
Die Ergebnisse können grob in zwei Gruppen unterteilt werden: In einigen Fällen besteht ein Großteil der Ergebnisliste aus nahezu identischen Audiodateien. Dies ist dadurch zu erklären, dass in D2 längere Audiodateien in einzelne 5 s lange Audiodateien unterteilt wurden. Bei manchen Suchanfragen wurden in der Ergebnisliste hauptsächlich Ausschnitte einer ursprünglichen Datei enthalten. In anderen Fällen enthalten die Ergebnislisten hauptsächlich Audiodateien, die von unterschiedlichen Aufnahmen stammen. Die Ähnlichkeiten innerhalb der Ergebnislisten mit unterschiedlichen Aufnahmen sind unterschiedlich stark ausgeprägt. Es ist in keinem Fall auszuschließen, dass es noch ähnlichere Klänge in D2 gegeben hätte, die fälschlicherweise nicht gefunden wurden.

[21] Siehe `https://github.com/ESchae/SimilarSoundSearch/tree/master/Evaluation/Evaluation_with_D2` (6. Juni 2016). Die Ergebnisse sind entsprechend der Reihenfolge in der Ergebnisliste nummeriert, gefolgt vom Dateinamen aus dem originalen Datenset (ESC-US, siehe Unterabschnitt 4.1.2). Das erste Ergebnis ist die jeweilige Suchanfrage.

4.5 Einschränkungen

Die Evaluation unterliegt einigen Einschränkungen, die hauptsächlich auf finanzielle und zeitliche Rahmenbedingungen der vorliegenden Arbeit zurückzuführen sind. Unter anderem wäre es interessant gewesen, folgende Aspekte weiter zu untersuchen:

Auswahl der Merkmale Die Auswahl der Merkmale wurde abgesehen vom Effekt der MFCC nicht evaluiert. Um die Auswahl zu verbessern, sollte einerseits die Korrelation der Merkmale untereinander überprüft werden [85]. Andererseits kann mithilfe von etablierten Methoden die Wahl der Merkmale unterstützt werden. Dadurch könnten noch viel mehr Merkmale in die Vorauswahl miteinbezogen werden. Aus den vielen Methoden der Merkmalsauswahl sollte die Methode selbst wiederum sorgsam ausgewählt werden. Je nach Anwendung kann es beispielsweise wünschenswert sein, wenn die Merkmale im Merkmalsvektor anschließend noch intuitiv im Bezug zur Wahrnehmung zu interpretieren sind. Außerdem sollten Audiomerkmale, die den selben psychoakustischen Parameter modellieren, miteinander verglichen werden. Insbesondere der Einfluss der MFCC und der Bezug zur Wahrnehmung müssten stärker untersucht werden.

Parameter Der Algorithmus enthält zahlreiche Parameter, die jeweils einzeln sowie in Kombination miteinander untersucht werden müssten. Beispielsweise wurde bei den Audiomerkmalen in der Regel die standardmäßige Einstellung der Parameter von Essentia übernommen. Auch der Einfluss von Framelänge und Hop-Size müsste untersucht werden.

Distanzmaß und Aggregation Der Einfluss des Distanzmaßes auf die Suchergebnisse müsste untersucht werden [wie beispielsweise in 98]. Das Gleiche gilt für die Generierung des Merkmalsvektors. Hier müssten die verwendeten Maße zur Aggregation evaluiert werden.

Gesamtkonzeption Die Evaluationsmaße könnten um weitere ergänzt werden, um ein differenzierteres Ergebnis zu erhalten. Außerdem wurde mithilfe von Crowdsourcing sozusagen *a priori* evaluiert, indem die paarweisen Distanzen zum Aufbau eines Benchmarks ermittelt wurden. Bei der tatsächlichen Suche sieht ein Nutzer hingegen eine Liste von Ergebnissen. Möglicherweise wird die Relevanz von Ergebnissen innerhalb einer Liste von Klängen anders beurteilt als paarweise Ähnlichkeiten. Im Idealfall sollten daher zusätzlich direkt die Suchergebnisse des Suchalgorithmus – also *a posteriori* – empirisch evaluiert werden.

5 Ausblick

Wie bereits in Unterabschnitt 3.1.3 erwähnt, gibt es vor allem zwei Aspekte, die in Zukunft für die inhaltsbezogene Suche nach gleichartigen Klängen verbessert werden könnten. Einerseits die stärkere Berücksichtigung von Kontexteffekten, was beispielsweise in Form von erweiterter Suchfunktionalität umgesetzt werden könnte. Andererseits muss der Bezug zur Wahrnehmung noch stärker ausgebaut werden. Beide Ansätze werden im Folgenden genauer beschrieben.

5.1 Erweiterte Suchfunktionen

Folgende Aspekte könnten in Zukunft im Hinblick auf die Suchfunktion verbessert werden:

Verfeinerte Auflösung Im vorliegenden Algorithmus werden Audiodateien immer als Ganzes verglichen. Audiodateien können verschiedene distinkte Klänge enthalten. In manchen Fällen ist es beabsichtigt, dieses Klanggemischt als Ganzes zu betrachten. Man spricht in dem Kontext auch von *Soundscapes* [der Begriff wurde eingeführt von Murray Schafer in 97]. Für manche Anwendungen sollten Audiosignale vor der Suche in einzelne Klänge unterteilt werden. Einzelne Klänge könnten wiederum in Unterabschnitte unterteilt werden. Dadurch könnten Suchergebnisse weiter verfeinert werden („Der Anfang von Klang A ist ähnlich der Mitte von Klang B“). Eine solche Verfeinerung der Suche wurde bereits in [118] teilweise umgesetzt. Die Unterteilung von Audiosignalen fällt in den Forschungsbereich *Audio Segmentation*.

Verschiedene Arten von Ähnlichkeit Wie in Unterabschnitt 2.5.1 verdeutlicht, gibt es verschiedene Arten von Ähnlichkeit. Es fällt nicht immer leicht, die rein akustische Ähnlichkeit von der semantischen Ähnlichkeit zu trennen. Dies bestätigte sich unter Anderem in den Crowdflower Ergebnissen, insbesondere bei den Testfragen. Möglicherweise entspricht der menschlichen Kognition eher eine Suchfunktion, in der akustische Ähnlichkeit und Ähnlichkeit der Klangquelle kombiniert werden [vgl. 46, S. 68] oder zumindest beides ermöglicht wird, da beide Ansätze je nach Anwendung Vor- und Nachteile besitzen können. Die Kombination dieser Ansätze wurde exemplarisch bereits in [51], [69] oder [71] umgesetzt.

Alternative Suchanfragen Die sprachliche Beschreibung von Klang ist in der Regel schwierig. Das query-by-example Paradigma hat den Vorteil, dass die Suchanfrage nicht verbalisiert werden muss. Voraussetzung dafür ist, dass man ein passendes Beispiel dafür hat, wonach man sucht. Esling & Agon fassen passend zusammen [Übersetzung durch Autorin; 26, S. 2058]:

> [E]in Hauptproblem bei der Suche nach Audiodaten (...) liegt in der Spezifikation der Suchanfrage an sich. (...) Audiodateien als Suchanfragen sind selbst komplex und facettenreich. Verschiedene Autoren (...) haben darauf hingewiesen, dass die meisten Nutzer zu Beginn der Suche nur eine vage Vorstellung davon haben, was sie suchen.

Die Autoren bieten Nutzern daher die Möglichkeit, bestimmte Klangaspekte zu zeichnen oder mit der Stimme zu imitieren. Tatsächlich scheint die menschliche Stimme, trotz Beschränkungen der Ausdrucksfähigkeit, dazu geeignet, entscheidende Audiomerkmale ausreichend repräsentieren zu können [58, 57]. Eine andere Möglichkeit wäre, Audiodaten auf intuitive Weise zu visualisieren, sodass das Durchsuchen auch ohne konkrete Vorstellung effizient gestaltet werden kann.[1]

Berücksichtigung von Kontexteffekten In Unterabschnitt 2.5.3 wurde verdeutlicht, dass es so etwas wie eine globale Ähnlichkeit nicht gibt. Je nach Suchanfrage können unterschiedliche Klangaspekte relevant sein. Das heißt, Subjektivität und Kontextabhänigkeit von Ähnlichkeit müssen mehr berücksichtigt werden. Zum Beispiel könnte dem Nutzer ermöglicht werden, bei der Suche explizit einen individuellen Schwerpunkt auf bestimmte Merkmale zu setzen [vgl. sub-similarities innerhalb von musikalischer Ähnlichkeit in 91]. Im vorliegenden Algorithmus und auch in der Freesound API ist es möglich anzugeben, welche Audiomerkmale für die Suche verwendet werden sollen. Um diese Funktion effektiv nutzen zu können, muss sich ein Nutzer mit Audiomerkmalen auskennen. Da jedesmal der Suchraum neu aufgebaut werden muss, ist sie auch nicht effizient. Wesentlich nutzerfreundlicher wären vorgegebene Beispielkategorien. Intuitive und dennoch aufschlussreiche Beispielkategorien zu finden ist wiederum nicht einfach. Die Evaluation wird dadurch auch aufwendiger.
Wie im vorherigen Abschnitt beschrieben, haben Nutzer oftmals nur eine vage Vorstellung davon was sie suchen. In diesem Fall könnte beispielsweise sogenanntes *Relevance Feedback* die individuelle Schwerpunktgestaltung eines Nutzers implizit unterstützen. Beim Relevance Feedback gibt der Nutzer an, welche Ergebnisse er relevant findet. Ausgehend davon wird anschließend nochmals eine verfeinerte Suche durchgeführt. Die Erweiterung um Relevance Feedback wäre relativ einfach. Einige Systeme verwenden dieses bereit [73, S. 77]. Alternativ könnte durch einen

[1] Ein Beispiel ist die Software Soundtorch 2.0 (`http://soundtorch.com/`, 5. Juni 2016), die leider nicht mehr betrieben zu werden scheint.

Algorithmus wie in [26] dem Nutzer anstatt einer globalen Ergebnisliste mehrere Untermengen von Ergebnislisten präsentiert werden, in denen jeweils unterschiedliche Aspekte berücksichtigt wurden.

5.2 Verstärkter Wahrnehmungsbezug

Um die Qualität der inhaltsbezogenen Suche nach gleichartigen Klängen in Zukunft zu verbessern, sind sicherlich elaboriertere Audiomerkmale nötig. In verschiedenen Studien wurden neue Audiomerkmale vorgeschlagen. Dazu gehören beispielsweise die Berücksichtigung von morphologischen Klangeigenschaften entsprechend der Klangontologie von Pierre Schaeffer [88, 31].
Die in der vorliegenden Arbeit extrahierten Audiomerkmale lassen fast alle eine psychoakustische Interpretation zu. Trotzdem basieren sie nicht auf psychoakustischen Modellen der auditiven Wahrnehmung. Dies gilt für die meisten Audiomerkmale [73, S. 103–108]. Audiomerkmale, die auf psychoakustischen Modellen beruhen, versuchen die menschliche Verarbeitung zu simulieren. Solche Audiomerkmale führen nicht automatisch zu besseren Ergebnissen [59, S. 77], denn nach wie vor sind viele Abläufe in der Verarbeitung ungeklärt und Modelle dadurch ungenau und fehlerbehaftet.
Trotzdem muss in Zukunft der Bezug zur Wahrnehmung stärker ausgebaut werden. Das grundlegende Problem in der Entwicklung neuer Audiomerkmale ist die sogenannte *semantische Lücke* [73, S. 75]. Es ist relativ einfach, neue Merkmale zu entwickeln. Der Bezug zur Wahrnehmung und folglich die Relevanz für einen Suchalgorithmus ist dagegen oft schwer zu beurteilen. Darüber hinaus ist davon auszugehen, dass eine bloße Verbesserung von Audiomerkmalen nicht ausreicht. Stattdessen muss die Gesamtkonzeption einen stärkeren Wahrnehmungsbezug erreichen. Aucouturier & Pachet kommen in einem Review über verschiedene Methoden zur Modellierung von Klangfarbenähnlichkeit zu folgender Schlussfolgerung, die auch auf die vorliegende Arbeit zu übertragen ist [Übersetzung durch Autorin; 79, S. 10]:

> Die Beschränkungen, die in dieser Arbeit festgestellt wurden (...), suggerieren, dass der übliche Weg zur Klangfarbenähnlichkeit möglicherweise nicht der Optimale ist. (...) Wir glauben, dass wesentliche Verbesserungen (...) nicht über weitere Variationen des üblichen Modells erreicht werden können, sondern dass ein tieferes Verständnis der kognitiven Prozesse nötig ist, die der Wahrnehmung von komplexen (...) Klangfarben sowie der Beurteilung deren Ähnlichkeit zugrunde liegen.

Auf ähnliche Weise fassen Siedenburg, Fujinaga & McAdams zusammen [Übersetzung durch Autorin; 99, S. 34]:

> Wahrnehmung und Computermodellierung sind jedoch nicht unabhängig voneinander, sobald die modellierten Phänomene schon an sich von psychologischer Natur sind. Dies wurde weitestgehend ignoriert in MIR,[2] ein Bereich, der genauso wie ein Großteil des Bereichs der Signalverarbeitung zögerlich ist, systematisch wahrnehmungsbasierte Evaluationen in seine Methoden zu integrieren, selbst wenn viele der Systeme Menschen als Zielgruppe ansprechen. (...) Trotz harter Arbeit an Audiomerkmalen (...) scheint sich die durch Precision und Recall[3] gemessene Leistung nicht signifikant verbessert zu haben. Dies könnte eine natürliche Konsequenz davon sein, die inhärente psychologische Natur von (...) Ähnlichkeit vernachlässigt zu haben.

Um die semantische Lücke weiter zu verringern, muss in Zukunft die interdisziplinäre Zusammenarbeit zwischen den Forschungsbereichen Audio Information Retrieval, Psychoakustik, Phänomenologie sowie Kognitions- und Neurowissenschaften gefördert werden [positive Beispiele jüngster Zeit finden sich unter Anderem in 81, 80, 81, oder auch 49]. Die interdisziplinäre Zusammenarbeit wird bislang vor allem durch unterschiedliche Methoden sowie Zielsetzungen und daraus resultierenden gegenseitigen Missverständnissen erschwert [sehr gut erörtert wird dies in 2, sowie 99]. Diese zu klären und zu überbrücken muss ein nächster konsequenter Schritt sein.

[2] Gemeint ist der Bereich *Music Information Retrieval*, hier synonym zum Audio Information Retrieval verwendet.

[3] Precision und Recall sind Standartevaluationsmaße aus dem Bereich Information Retrieval.

A Anhang

A.1 Multidimensionale Skalierung

Es gibt verschiedene Varianten von MDS. Meistens wird eine Verteilung im euklidischen Raum angenommen (zum euklidischen Abstand siehe Unterabschnitt 3.4.1). Andere Varianten von MDS erweitern das klassische euklidische Distanzmodell. Beispielsweise kann das Modell um sogenannte Spezifitäten (engl. specificities) erweitern werden. Dies sind spezifische Merkmale, die nur bestimmte Klangfarben besitzen und die sie von allen anderen Klangfarben unterscheiden. Spezifizitäten sind oft schwer zu interpretieren. Ebenso kann berücksichtigt werden, dass Versuchspersonen bei der Beurteilung der Ähnlichkeit einzelnen Dimensionen möglicherweise unterschiedlich stark einbeziehen. Um Parameter zu reduzieren, können die Versuchspersonen anhand ihrer Wahrnehmung und Beurteilungsstrategie in latente Klassen unterteilt werden. Eine dieser Erweiterungen des klassischen euklidischen Distanzmodells ist das sogenannte CLASCAL Modell, das wie folgt definiert ist [64, S. 37]:

$$d_{ijt} = \left[\sum_{r=1}^{R} w_{tr}(x_{ir} - x_{jr})^2 + v_t(s_i + s_j) \right]^{\frac{1}{2}} . \tag{A.1}$$

d_{ijt} ist in diesem Modell der Abstand zwischen zwei Elementen i und j innerhalb der latenten Klasse t, x_{ir} und x_{jr} die Koordinaten der Elemente i beziehungsweise j in Dimension r, w_{tr} ist das Gewicht, das Dimension r durch Klasse t zugesprochen wird, s_i ist die Spezifizität von Element i und v_t ist das Gewicht, dass allen Spezifizitäten durch Klasse t zugesprochen wird.

A.2 R-Test detaillierte Ergebnisse

Tabelle A.1: p-Werte (in %) des R-Tests zwischen den Algorithmen Ohne MFCC (Ohne), Nur MFCC (Nur), Alle Merkmale (Alle) und Freesound (Free) für die verwendeten Evaluationsmaße bei linearer sowie logarithmischer Skalierung.

Genauigkeit ($\delta = 0$, lin.)

p (%)	Ohne	Nur	Alle	Free
Ohne	·	49.61	86.13	**0.20**
Nur	49.61	·	25.00	**0.20**
Alle	86.13	25.00	·	**0.20**
Free	**0.20**	**0.20**	**0.20**	·

Genauigkeit ($\delta = 0$, log.)

p (%)	Ohne	Nur	Alle	Free
Ohne	·	**0.20**	**0.59**	**0.20**
Nur	**0.20**	·	75.39	**0.78**
Alle	**0.59**	75.39	·	**0.78**
Free	**0.20**	**0.78**	**0.78**	·

Genauigkeit ($\delta = 1$, lin.)

p (%)	Ohne	Nur	Alle	Free
Ohne	·	**0.20**	**0.39**	**0.20**
Nur	**0.20**	·	55.47	**0.39**
Alle	**0.39**	55.47	·	**0.20**
Free	**0.20**	**0.39**	**0.20**	·

Genauigkeit ($\delta = 1$, log.)

p (%)	Ohne	Nur	Alle	Free
Ohne	·	7.81	**1.17**	**1.17**
Nur	7.81	·	**3.91**	**2.73**
Alle	**1.17**	**3.91**	·	6.25
Free	**1.17**	**2.73**	6.25	·

Genauigkeit ($\delta = 2$, lin.)

p (%)	Ohne	Nur	Alle	Free
Ohne	·	**0.20**	**0.39**	**0.20**
Nur	**0.20**	·	67.19	**0.78**
Alle	**0.39**	67.19	·	**0.78**
Free	**0.20**	**0.78**	**0.78**	·

Genauigkeit ($\delta = 2$, log.)

p (%)	Ohne	Nur	Alle	Free
Ohne	·	79.30	77.93	81.25
Nur	79.30	·	100.0	95.31
Alle	77.93	100.0	·	100.0
Free	81.25	95.31	100.0	·

Dur. Abweichung (lin.)

p (%)	Ohne	Nur	Alle	Free
Ohne	·	39.84	16.80	**3.12**
Nur	39.84	·	37.89	47.66
Alle	16.80	37.89	·	74.22
Free	**3.12**	47.66	74.22	·

Dur. Abweichung (log.)

p (%)	Ohne	Nur	Alle	Free
Ohne	·	79.88	86.72	63.28
Nur	79.88	·	87.89	89.84
Alle	86.72	87.89	·	73.44
Free	63.28	89.84	73.44	·

nDCG (lin.)

p (%)	Ohne	Nur	Alle	Free
Ohne	·	70.90	**3.32**	48.05
Nur	70.90	·	12.70	76.17
Alle	**3.32**	12.70	·	31.25
Free	48.05	76.17	31.25	·

nDCG (log.)

p (%)	Ohne	Nur	Alle	Free
Ohne	·	70.90	**3.32**	48.05
Nur	70.90	·	12.70	76.17
Alle	**3.32**	12.70	·	31.25
Free	48.05	76.17	31.25	·

Tabelle A.2: Übersicht über die 150 Klänge aus D1. In den ersten zehn Reihen stehen die zehn Beispielklänge, zu denen jeweils Cluster von 14 ähnlichen Klängen folgen. Der jeweilige Cluster steht in der zweiten Spalte. In der dritten Spalte steht eine kurze Beschreibung des Klangs. Diese ist sehr knapp gehalten und kann nicht das Hören des Klangs ersetzen. In der vierten Spalte steht die ID des Sounds auf Freesound. In der fünften Spalte steht der Name der Datei, die auf `https://github.com/ESchae/SimilarSoundSearch/tree/master/Evaluation/D1/` (6. Juni 2016) im Ordner audiofiles angehört werden kann. Der Name setzt sich zusammen aus Freesound ID des Beispielklangs des Clusters, Distanz zwischen Beispielklang und Klang auf Freesound sowie der Freesound ID des Klangs. Innerhalb eines Clusters sind die Klänge nach aufsteigender Distanz sortiert.

Nr.	Cluster	Beschreibung	ID	Filename
1	1 - Katze	Ausgedehntes Miauen einer Katze	110011	0-110011.mp3
2	2 - Hydrant	Zischen eines undichten Hydranten	2155	0-2155.mp3
3	3 - Tür	Quietschende Tür (erst d‴/dis‴ - dann e‴)	22362	0-22362.mp3
4	4 - Spielzeug	Zweimaliges Quietschen von Gummispielzeug (c‴-dis‴)	240015	0-240015.mp3
5	5 - Uhr	Gleichmäßiges Ticken einer Uhr	264498	0-264498.mp3
6	6 - Lachen	Synthetisches und schnell abgespieltes Lachen	325462	0-325462.mp3
7	7 - Regen	Gleichmäßiges Rauschen von Regen	337791	0-337791.mp3
8	8 - Grollen	Synthetisches tiefes und einmaliges gehauchtes Grollen	50802	0-50802.mp3
9	9 - Feuerwerk	Feuerwerk mit mehrfachen Explosionen und Raketen	5560	0-5560.mp3
10	10 - Schrei	Mehrere sehr hohe Schreie eines Mädchens (d⁗)	82402	0-82402.mp3
11	1 - Katze	Katze (ohne wahrnehmbaren Unterschied zu 110011)	256452	110011-0.98-256452.mp3
12	1 - Katze	Katze (ähnlich 62215 und 62216 und 62209)	62213	110011-1.11-62213.mp3
13	1 - Katze	Quängelndes Kind	315829	110011-1.39-315829.mp3
14	1 - Katze	Katze (ähnlich 62215 und 62213 und 62209)	62216	110011-1.39-62216.mp3
15	1 - Katze	Katze (eher tief und meckernd)	156643	110011-1.40-156643.mp3
16	1 - Katze	Synthetische Klänge ähnlich verzerrter E-Gitarre	344872	110011-1.41-344872.mp3
17	1 - Katze	Katze (ähnlich 62216 und 62213 und 62215)	62209	110011-1.41-62209.mp3
18	1 - Katze	Szene mit Lautsprecher und Menschenstimmen	41251	110011-1.42-41251.mp3
19	1 - Katze	Synthetische Frauenstimme ('Activating alarm system')	169206	110011-1.43-169206.mp3
20	1 - Katze	Katze (ähnlich 62216 und 62213 und 62209)	62215	110011-1.43-62215.mp3
21	1 - Katze	Quängelnde Kinder	315834	110011-1.45-315834.mp3
22	1 - Katze	Szene Flugzeugdurchsagen	339894	110011-1.45-339894.mp3

23	1 - Katze	Wiederholte Lautsprecherdurchsagen	202192	110011-1.48-202192.mp3
24	1 - Katze	Synthetischer Klang mit Glissando	133542	110011-1.49-133542.mp3
25	2 - Hydrant	Laufendes Wasser in Dusche	140391	2155-0.54-140391.mp3
26	2 - Hydrant	Brutzelndes Fett in Pfanne	165116	2155-0.60-165116.mp3
27	2 - Hydrant	Laufendes Wasser	135003	2155-0.61-135003.mp3
28	2 - Hydrant	Zischen beim Kochen von Wasser im Teekessel	190936	2155-0.65-190936.mp3
29	2 - Hydrant	Fahrradkette im Leerlauf	327542	2155-0.65-327542.mp3
30	2 - Hydrant	Abbrennender Docht einer Bombe	329045	2155-0.65-329045.mp3
31	2 - Hydrant	Regen	167206	2155-0.66-167206.mp3
32	2 - Hydrant	In Abwasserkanal strömender Regen	210827	2155-0.66-210827.mp3
33	2 - Hydrant	Regen	326445	2155-0.66-326445.mp3
34	2 - Hydrant	Zirpen von Grillen	321444	2155-0.67-321444.mp3
35	2 - Hydrant	Beatbox Hi-Hat	70626	2155-0.67-70626.mp3
36	2 - Hydrant	Schleifendes Geräusch	201635	2155-0.69-201635.mp3
37	2 - Hydrant	Laufender Wasserhahn	202529	2155-0.69-202529.mp3
38	2 - Hydrant	Verrauschtes Lagerfeuer	241318	2155-0.69-241318.mp3
39	3 - Tür	Mehrfach quietschende Tür (immer wieder dis‴- gis')	174067	22362-1.11-174067.mp3
40	3 - Tür	Geige (dis‴)	247658	22362-1.13-247658.mp3
41	3 - Tür	Klarinette (e‴/f‴)	248927	22362-1.15-248927.mp3
42	3 - Tür	Klarinette (e‴)	249016	22362-1.15-249016.mp3
43	3 - Tür	Geige (dis‴)	247842	22362-1.16-247842.mp3
44	3 - Tür	Geige (dis‴)	250953	22362-1.16-250953.mp3
45	3 - Tür	Klarinette (dis‴)	249015	22362-1.19-249015.mp3
46	3 - Tür	Klarinette (dis‴)	248926	22362-1.22-248926.mp3
47	3 - Tür	Nacheinander angeschlagene Gläser (f″-e″-g″-dis″)	69552	22362-1.27-69552.mp3
48	3 - Tür	Geige (dis‴)	247459	22362-1.28-247459.mp3
49	3 - Tür	Klarinette (d‴/dis″)	248609	22362-1.29-248609.mp3
50	3 - Tür	Geige (d‴)	247456	22362-1.30-247456.mp3
51	3 - Tür	Geige (dis‴)	247731	22362-1.30-247731.mp3
52	3 - Tür	Geige (dis‴)	247880	22362-1.30-247880.mp3
53	4 - Spielzeug	Quietschende Tür (mehrere Töne um d‴ herum)	266319	240015-1.51-266319.mp3
54	4 - Spielzeug	Synthetische Töne	334805	240015-1.59-334805.mp3
55	4 - Spielzeug	Schnelles Trommeln auf Metallbüchsen	124716	240015-1.63-124716.mp3
56	4 - Spielzeug	Klopfen gegen Glas (d‴/dis‴)	118560	240015-1.66-118560.mp3

57	4 - Spielzeug	Synthetische Frauenstimme ('sugoisugoi')	278141	240015-1.71-278141.mp3
58	4 - Spielzeug	Ohne Wahrnehmbaren Unterschied zu 118560	118561	240015-1.74-118561.mp3
59	4 - Spielzeug	Synthetischer Klang - Messerwurf	181679	240015-1.74-181679.mp3
60	4 - Spielzeug	Synthetische Klänge	296486	240015-1.75-296486.mp3
61	4 - Spielzeug	Quietschen beim Wischen über Glasplatte	245193	240015-1.76-245193.mp3
62	4 - Spielzeug	Synthetischer Ton - kurzes Bellen	216994	240015-1.77-216994.mp3
63	4 - Spielzeug	Loop mit verschiedenen Perkussionsinstrumenten	223676	240015-1.77-223676.mp3
64	4 - Spielzeug	Trommeln gegen Keramikglocke	254755	240015-1.77-254755.mp3
65	4 - Spielzeug	Quietschende Tür (Glissando c‴-a")	70377	240015-1.77-70377.mp3
66	4 - Spielzeug	Mehrmaliges Zupfen von Harfe (a")	210178	240015-1.78-210178.mp3
67	5 - Uhr	Rhythmisches Surren von Kamera Winder	125319	264498-0.63-125319.mp3
68	5 - Uhr	Blindenstock auf verschiedenen Untergründen	174604	264498-0.71-174604.mp3
69	5 - Uhr	Rhythmisches Schnipsen (ähnlich 174251)	174252	264498-0.75-174252.mp3
70	5 - Uhr	Tippen auf Tastatur	240280	264498-0.76-240280.mp3
71	5 - Uhr	Rhythmisches Schnipsen (ähnlich 174148)	174353	264498-0.77-174353.mp3
72	5 - Uhr	Münzen beim Einwurf in Automat	185045	264498-0.77-185045.mp3
73	5 - Uhr	Knistern mit Alufolie	264456	264498-0.77-264456.mp3
74	5 - Uhr	Rhythmisches Schnipsen (ähnlich 174252)	174251	264498-0.78-174251.mp3
75	5 - Uhr	Tippen auf Tastatur	186134	264498-0.79-186134.mp3
76	5 - Uhr	Tickende Uhr	30608	264498-0.79-30608.mp3
77	5 - Uhr	Rhythmisches Schnipsen (ähnlich 17353)	174148	264498-0.80-174148.mp3
78	5 - Uhr	Tippen auf Tastatur	179385	264498-0.80-179385.mp3
79	5 - Uhr	Umblättern von Seiten	248045	264498-0.80-248045.mp3
80	5 - Uhr	Einsortieren von Gabeln in Behälter	344633	264498-0.80-344633.mp3
81	6 - Lachen	Jammerndes Kind	344039	325462-1.45-344039.mp3
82	6 - Lachen	Husten	251489	325462-1.53-251489.mp3
83	6 - Lachen	Synthetisches musikalisches Pattern	40728	325462-1.57-40728.mp3
84	6 - Lachen	Mehrmaliges Lachen von Frau	37241	325462-1.58-37241.mp3
85	6 - Lachen	Schnell abgespieltes Lachen oder Stimme	89467	325462-1.59-89467.mp3
86	6 - Lachen	Synthetisches musikalisches Pattern	200837	325462-1.62-200837.mp3
87	6 - Lachen	Kichern	119450	325462-1.64-119450.mp3
88	6 - Lachen	Synthetische Loop	316793	325462-1.64-316793.mp3
89	6 - Lachen	Lachen von Frau	319346	325462-1.64-319346.mp3
90	6 - Lachen	Mehrmaliges Lachen von Kind	4237	325462-1.64-4237.mp3

91	6 - Lachen	Mehrmaliges Lachen von Kind	55209	325462-1.65-55209.mp3
92	6 - Lachen	Synthetische Jingle	7285	325462-1.65-7285.mp3
93	6 - Lachen	Lachen von Mann	19158	325462-1.66-19158.mp3
94	6 - Lachen	Lachen von Kind	45129	325462-1.66-45129.mp3
95	7 - Regen	Regen	213012	337791-0.42-213012.mp3
96	7 - Regen	Regen	326441	337791-0.48-326441.mp3
97	7 - Regen	Springbrunnen	79367	337791-0.48-79367.mp3
98	7 - Regen	Rauschen	197202	337791-0.51-197202.mp3
99	7 - Regen	Regen	110612	337791-0.52-110612.mp3
100	7 - Regen	Wasserfall	130231	337791-0.53-130231.mp3
101	7 - Regen	Fluss	191876	337791-0.54-191876.mp3
102	7 - Regen	Wasserfall	80834	337791-0.55-80834.mp3
103	7 - Regen	Rauschen	44695	337791-0.57-44695.mp3
104	7 - Regen	Kurzer Störsound	44734	337791-0.57-44734.mp3
105	7 - Regen	Synthetischer Klang - ähnlich Inlineskates	209650	337791-0.58-209650.mp3
106	7 - Regen	Springbrunnen	93985	337791-0.58-93985.mp3
107	7 - Regen	Meer Brandung	163615	337791-0.59-163615.mp3
108	7 - Regen	Regen	276928	337791-0.59-276928.mp3
109	8 - Grollen	Tiefes Grollen	66574	50802-0.79-66574.mp3
110	8 - Grollen	Lange Szene - Gruselstimmung	167277	50802-0.80-167277.mp3
111	8 - Grollen	Kurze Verkehrsszene	156091	50802-0.82-156091.mp3
112	8 - Grollen	Lange düstere Atmosphärenmusik	170989	50802-0.82-170989.mp3
113	8 - Grollen	Szene auf Straße mit vielen Menschen	343282	50802-0.82-343282.mp3
114	8 - Grollen	Kleinkind in Auto	74408	50802-0.82-74408.mp3
115	8 - Grollen	Verschiedene Szenen mit Menschen und Musik	178925	50802-0.83-178925.mp3
116	8 - Grollen	Tiefe Männerstimme erzählend mit viel Hall	240734	50802-0.83-240734.mp3
117	8 - Grollen	Stuhl über Boden gezogen bearbeitet	71153	50802-0.83-71153.mp3
118	8 - Grollen	Synthetischer Klang - ähnlich Raketenstart	116831	50802-0.84-116831.mp3
119	8 - Grollen	Tiefes Grollen	163445	50802-0.84-163445.mp3
120	8 - Grollen	Vorbeifahrender Zug	83930	50802-0.84-83930.mp3
121	8 - Grollen	Kurzer synthetischer Klang	83946	50802-0.84-83946.mp3
122	8 - Grollen	Synthetischer Gewitterschlag	84521	50802-0.85-84521.mp3
123	9 - Feuerwerk	Kanonenschüsse	239135	5560-0.80-239135.mp3
124	9 - Feuerwerk	Donner	30303	5560-0.83-30303.mp3

125	9 - Feuerwerk	Kurze synthetische Explosion	207322	5560-0.88-207322.mp3
126	9 - Feuerwerk	Synthetisches Feuer	17786	5560-0.94-17786.mp3
127	9 - Feuerwerk	Donner	199944	5560-0.94-199944.mp3
128	9 - Feuerwerk	Herunterfallendes Mikrofon	255209	5560-0.94-255209.mp3
129	9 - Feuerwerk	Auf Tische trommelnde Hände	81192	5560-0.94-81192.mp3
130	9 - Feuerwerk	Szene in Flugzeug - Sicherheitsanweisung bis Start	156845	5560-0.97-156845.mp3
131	9 - Feuerwerk	Musikalisches Pattern	321026	5560-0.99-321026.mp3
132	9 - Feuerwerk	Vorbeifliegendes Düsenflugzeug	189446	5560-1.02-189446.mp3
133	9 - Feuerwerk	Synthetische Klänge	61680	5560-1.04-61680.mp3
134	9 - Feuerwerk	Donner	214052	5560-1.06-214052.mp3
135	9 - Feuerwerk	Stuhl über Boden gezogen	143951	5560-1.07-143951.mp3
136	9 - Feuerwerk	Synthetische Loop - wie rhythmisches Atmen	61967	5560-1.09-61967.mp3
137	10 - Schrei	Drei hohe Mädchenschreie (absteigend d′′′′-ais′′′)	326332	82402-0.93-326332.mp3
138	10 - Schrei	Drei hohe Mädchenschreie (cis′′′′-cis′′′-e′′′)	242009	82402-1.08-242009.mp3
139	10 - Schrei	Klarinette (ais′′′)	248875	82402-1.10-248875.mp3
140	10 - Schrei	Melodischer Pfiff (c′′′′-f′′′′-h′′′)	66550	82402-1.11-66550.mp3
141	10 - Schrei	Bussardschreie (Glissando dis′′′′-c′′′′)	69504	82402-1.12-69504.mp3
142	10 - Schrei	Harfe (h′′′)	302793	82402-1.13-302793.mp3
143	10 - Schrei	Klarinette (ais′′′)	248660	82402-1.14-248660.mp3
144	10 - Schrei	Mikrowelle - erst Brummen dann Piepen (h′′′)	175395	82402-1.17-175395.mp3
145	10 - Schrei	Querflöte (c′′′′)	246989	82402-1.21-246989.mp3
146	10 - Schrei	Schlag gegen Glas (c′′′′/d′′′′)	321284	82402-1.21-321284.mp3
147	10 - Schrei	Glöckchen (c′′′′)	339820	82402-1.21-339820.mp3
148	10 - Schrei	Schlag gegen Glas (ais′/ais′′ und mehr)	42909	82402-1.22-42909.mp3
149	10 - Schrei	Melodischer Pfiff (dis′′′′-f′′′′-d′′′′)	66545	82402-1.23-66545.mp3
150	10 - Schrei	Schrei (gis′′-h′′-ais′′)	82418	82402-1.23-82418.mp3

Literatur

[1] Aldrich, K. M., Hellier, E. J. und Edworthy, J. “What determines auditory similarity? The effect of stimulus group and methodology”. In: *The Quarterly Journal of Experimental Psychology* 62.1 (2009), S. 63–83 (siehe S. 20, 29, 31, 53).

[2] Aucouturier, J.-J. und Bigand, E. “Seven problems that keep MIR from attracting the interest of cognition and neuroscience”. In: *Journal of Intelligent Information Systems* 41.3 (2013), S. 483–497 (siehe S. 76).

[3] Bailey, S. V. und Rice, S. V. “A Web Search Engine for Sound Effects”. In: *Audio Engineering Society Convention 119*. Audio Engineering Society, 2005 (siehe S. 5).

[4] Bast, H., Buchhold, B. und Haussmann, E. “Relevance Scores for Triples from Type-Like Relations”. In: ACM Press, 2015, S. 243–252 (siehe S. 63, 65).

[5] Berland, A. u. a. “Perception of Everyday Sounds: A Developmental Study of a Free Sorting Task”. In: *PLoS ONE* 10.2 (2015), S. 1–21 (siehe S. 31).

[6] Blum, T. u. a. “Method and article of manufacture for content-based analysis, storage, retrieval, and segmentation of audio information”. 5.918.223. 1999 (siehe S. 33).

[7] Bogdanov, D. u. a. “Essentia: An Audio Analysis Library for Music Information Retrieval.” In: *International Society for Music Information Retrieval Conference (ISMIR'13)*. Citeseer, 2013, S. 493–498 (siehe S. 34).

[8] Bonebright, T. L. “Perceptual structure of everyday sounds: A multidimensional scaling approach”. In: *Proceedings of the 7th International Conference on Auditory Display (ICAD2001)*. International Community for Auditory Display, 2001, S. 73–78 (siehe S. 20, 29).

[9] Bregman, A. S. *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press, 1994 (siehe S. 18).

[10] Brossier, P. M. “Automatic annotation of musical audio for interactive applications”. Diss. Queen Mary, University of London, 2006 (siehe S. 37, 39).

[11] Burgoyne, J. A. und McAdams, S. “A Meta-analysis of Timbre Perception Using Nonlinear Extensions to CLASCAL”. In: *Computer Music Modeling and Retrieval. Sense of Sounds*. Hrsg. von Kronland-Martinet, R., Ystad, S. und Jensen, K. Lecture Notes in Computer Science 4969. Springer Berlin Heidelberg, 2007, S. 181–202 (siehe S. 19, 20).

[12] Caclin, A. u. a. “Acoustic correlates of timbre space dimensions: a confirmatory study using synthetic tones”. In: *The Journal of the Acoustical Society of America* 118.1 (2005), S. 471–482 (siehe S. 19, 20, 31).

[13] Cambouropoulos, E. “How similar is similar?” In: *Musicae Scientiae* 13.1 suppl (2009), S. 7–24 (siehe S. 27).

[14] Cano, P. u. a. “A review of algorithms for audio fingerprinting”. In: *2002 IEEE Workshop on Multimedia Signal Processing*. 2002, S. 169–173 (siehe S. 4, 32, 33).

[15] Cartwright, M. B. und Pardo, B. “Novelty measures as cues for temporal salience in audio similarity”. In: *Proceedings of the second international ACM workshop on Music information retrieval with user-centered and multimodal strategies*. ACM Press, 2012, S. 51–56 (siehe S. 33, 52).

[16] Casey, M. u. a. “Content-Based Music Information Retrieval: Current Directions and Future Challenges”. In: *Proceedings of the IEEE* 96.4 (2008), S. 668–696 (siehe S. 4).

[17] Chachada, S. und Kuo, C.-C. “Environmental sound recognition: A survey”. In: *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2013 Asia-Pacific*. 2013, S. 1–9 (siehe S. 4).

[18] Clough, P. u. a. “Examining the limits of crowdsourcing for relevance assessment”. In: *Internet Computing, IEEE* 17.4 (2013), S. 32–38 (siehe S. 61).

[19] Cowling, M. und Sitte, R. “Comparison of techniques for environmental sound recognition”. In: *Pattern Recognition Letters* 24.15 (2003), S. 2895–2907 (siehe S. 4).

[20] de Cheveigné, A. und Kawahara, H. “YIN, a fundamental frequency estimator for speech and music”. In: *The Journal of the Acoustical Society of America* 111.4 (2002), S. 1917 (siehe S. 37, 39).

[21] Decock, L. und Douven, I. “Similarity After Goodman”. In: *Review of Philosophy and Psychology* 2.1 (2011), S. 61–75 (siehe S. 31).

[22] Dickerson, K., Gaston, J. R. und McCarty-Gibson, S. *Parameterizing Sound: Design Considerations for an Environmental Sound Database*. Techn. Ber. DTIC Document, 2015 (siehe S. 5, 53).

[23] Duan, S. u. a. “A Survey of Tagging Techniques for Music, Speech and Environmental Sound”. In: *Artificial Intelligence Review* 42.4 (2014), S. 637–661 (siehe S. 4).

[24] Ellermeier, W. und Hellbrück, J. “Hören – Psychoakustik – Audiologie”. In: *Handbuch der Audiotechnik*. Hrsg. von Weinzierl, S. VDI-Buch. Springer Berlin Heidelberg, 2008, S. 41–85 (siehe S. 13).

[25] Elliott, T. M., Hamilton, L. S. und Theunissen, F. E. “Acoustic structure of the five perceptual dimensions of timbre in orchestral instrument tones”. In: *The Journal of the Acoustical Society of America* 133.1 (2013), S. 389–404 (siehe S. 20).

[26] Esling, P. und Agon, C. “Multiobjective Time Series Matching for Audio Classification and Retrieval”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 21.10 (2013), S. 2057–2072 (siehe S. 33, 74, 75).

[27] Fastl, H. und Zwicker, E. *Psychoacoustics*. Springer Berlin Heidelberg, 2007 (siehe S. 13, 15, 16, 18, 19).

[28] Ferrer, R. und Eerola, T. “Semantic structures of timbre emerging from social and acoustic descriptions of music”. In: *EURASIP Journal on Audio, Speech, and Music Processing* 2011.1 (2011), S. 1–16 (siehe S. 3).

[29] Flückiger, B. *Sound Design: Die virtuelle Klangwelt des Films*. unveränderter Nachdruck der 3. Auflage 2007. Marburg: Schüren Verlag GmbH, 2006 (siehe S. 3, 18).

[30] Font, F., Roma, G. und Serra, X. “Freesound technical demo”. In: *Proceedings of the 21st ACM international conference on Multimedia*. ACM Press, 2013, S. 411–412 (siehe S. 2, 3, 5, 34).

[31] Gasser, M., Flexer, A. und Grill, T. “On computing Morphological Similarity of Audio Signals”. In: *Proceedings of the 8th Sound and Music Computing Conference*. 2011 (siehe S. 75).

[32] Gaver, W. W. “What in the world do we hear? an ecological approach to auditory event perception”. In: *Ecological Psychology* 5 (1993), S. 1–29 (siehe S. 27).

[33] Giordano, B. L., McAdams, S. und McDonnell, J. “Acoustical and Conceptual Information for the Perception of Animate and Inanimate Sound Sources”. In: *International Conference on Auditory Display*. Georgia Institute of Technology, 2007, S. 173–180 (siehe S. 20, 27, 28).

[34] Giordano, B. L. u. a. “Comparison of Methods for Collecting and Modeling Dissimilarity Data: Applications to Complex Sound Stimuli”. In: *Multivariate Behavioral Research* 46.5 (2011), S. 779–811 (siehe S. 28).

[35] Goldstone, R. L. “The role of similarity in categorization: providing a groundwork”. In: *Cognition* 52.2 (1994), S. 125–157 (siehe S. 28).

[36] Gregg, M. K. und Samuel, A. G. “The importance of semantics in auditory representations”. In: *Attention, Perception, & Psychophysics* 71.3 (2009), S. 607–619 (siehe S. 27).

[37] Guastavino, C. “Categorization of environmental sounds.” In: *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale* 61.1 (2007), S. 54–63 (siehe S. 27).

[38] Gygi, B. “Factors in the identification of environmental sounds”. Diss. Indiana University, 2001 (siehe S. 7, 28).

[39] Gygi, B., Kidd, G. R. und Watson, C. S. “Similarity and categorization of environmental sounds”. In: *Perception & Psychophysics* 69.6 (2007), S. 839–855 (siehe S. 4, 20, 27–29).

[40] Gygi, B., Kidd, G. R. und Watson, C. S. “Spectral-temporal factors in the identification of environmental sounds”. In: *The Journal of the Acoustical Society of America* 115.3 (2004), S. 1252 (siehe S. 28).

[41] Gygi, B. und Shafiro, V. “Development of the Database for Environmental Sound Research and Application (DESRA): Design, Functionality, and Retrieval Considerations”. In: *EURASIP Journal on Audio, Speech, and Music Processing* 2010 (2010), S. 1–12 (siehe S. 53).

[42] Haitsma, J. und Kalker, T. “A Highly Robust Audio Fingerprinting System With an Efficient Search Strategy”. In: *Journal of New Music Research* 32.2 (2003), S. 211–221 (siehe S. 4).

[43] Hall, D. E. *Musikalische Akustik: ein Handbuch*. Schott, 2008 (siehe S. 8–10, 13, 15–19).

[44] Helén, M. und Virtanen, T. “Audio Query by Example Using Similarity Measures between Probability Density Functions of Features”. In: *EURASIP Journal on Audio, Speech, and Music Processing* 2010 (2010), S. 1–12 (siehe S. 25, 33, 52).

[45] Hocking, J. u. a. “NESSTI: Norms for Environmental Sound Stimuli”. In: *PLoS ONE* 8.9 (2013), S. 1–12 (siehe S. 5, 28).

[46] Houix, O. u. a. “A lexical analysis of environmental sound categories.” In: *Journal of Experimental Psychology: Applied* 18.1 (2012), S. 52–80 (siehe S. 7, 20, 27, 73).

[47] Jadhav, S. M. und Patil, V. S. “Review of significant researches on multimedia information retrieval”. In: *2012 International Conference on Communication, Information Computing Technology (ICCICT)*. 2012, S. 1–6 (siehe S. 3).

[48] Järvelin, K. und Kekäläinen, J. “IR evaluation methods for retrieving highly relevant documents”. In: *Proceedings of the 23th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 2000, S. 41–48 (siehe S. 65).

[49] Ji, X. u. a. “Analysis of music/speech via integration of audio content and functional brain response”. In: *Information Sciences* 297 (2015), S. 271–282 (siehe S. 76).

[50] Kim, H.-G., Moreau, N. und Sikora, T. *MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval*. John Wiley & Sons, 2005 (siehe S. 4).

[51] Knees, P. u. a. “A Music Search Engine Built Upon Audio-based and Web-based Similarity Measures”. In: *Proceedings of the 30th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 2007, S. 447–454 (siehe S. 73).

[52] Krumhansl, C. L. “Why is Musical Timbre so hard to understand?” In: *Structure and Perception of Electroacoustic Sound and Music*. Hrsg. von Nielzen, S. und Olsson, O. Amsterdam: Elsevier, 1989 (siehe S. 18).

[53] Labuschagne, I. B. und Hanekom, J. J. “Preparation of stimuli for timbre perception studies”. In: *The Journal of the Acoustical Society of America* 134.3 (2013), S. 2256–2267 (siehe S. 19).

[54] Lallemand, I., Schwarz, D. und Artieres, T. “Content-based retrieval of environmental sounds by multiresolution analysis”. In: *SMC2012*. 2012, S. 1–8 (siehe S. 2, 33, 52).

[55] Lease, M. und Yilmaz, E. “Crowdsourcing for information retrieval”. In: *ACM SIGIR Forum*. ACM Press, 2012, S. 66–75 (siehe S. 54).

[56] Lemaitre, G. u. a. “Listener expertise and sound identification influence the categorization of environmental sounds.” In: *Journal of Experimental Psychology: Applied* 16.1 (2010), S. 16–32 (siehe S. 27, 31).

[57] Lemaitre, G. u. a. “Vocal imitations and the identification of sound events”. In: *Ecological psychology* 23.4 (2011), S. 267–307 (siehe S. 74).

[58] Lemaitre, G. u. a. “Vocal imitations of basic auditory features”. In: *The Journal of the Acoustical Society of America* 139.1 (2016), S. 290–300 (siehe S. 74).

[59] Lerch, A. *An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics*. Wiley-IEEE Press, 2012 (siehe S. 12, 15, 20–24, 34, 37, 41, 42, 44, 45, 47, 48, 50, 75).

[60] Lew, M. S. "Content-Based Multimedia Information Retrieval: State of the Art and Challenges". In: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 2.1 (2006), S. 1–19 (siehe S. 3).

[61] Logan, B. und Salomon, A. "A music similarity function based on signal analysis". In: *Multimedia and Expo, 2001. ICME 2001. IEEE International Conference on*. IEEE, 2001, S. 745–748 (siehe S. 4).

[62] Malik, H. *Content Based Audio Indexing and Retrieval: An Overview*. Technischer Bericht. 2003 (siehe S. 2).

[63] Marcell, M. M. u. a. "Confrontation naming of environmental sounds". In: *Journal of Clinical and Experimental Neuropsychology* 22.6 (2000), S. 830–864 (siehe S. 27).

[64] McAdams, S. "Musical Timbre Perception". In: *The Psychology of Music*. Elsevier, 2013, S. 35–67 (siehe S. 18–20, 77).

[65] McAdams, S., Beauchamp, J. W. und Meneguzzi, S. "Discrimination of musical instrument sounds resynthesized with simplified spectrotemporal parameters". In: *The Journal of the Acoustical Society of America* 105.2 (1999), S. 882–897 (siehe S. 31).

[66] McAdams, S. u. a. "A meta-analysis of acoustic correlates of timbre dimensions". In: *The Journal of the Acoustical Society of America* 120.5 (2006), S. 3275–3276 (siehe S. 20, 43).

[67] McAdams, S. u. a. "Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes". In: *Psychological Research* 58.3 (1995), S. 177–192 (siehe S. 19).

[68] Mechtley, B., Cook, P. und Spanias, A. "Sound Mapping on the Web: Current Solutions and Future Directions". In: *Proceedings of the Symposium on Acoustic Ecology*. 2013, S. 1–4 (siehe S. 2).

[69] Mechtley, B. u. a. "Combining semantic, social, and acoustic similarity for retrieval of environmental sounds". In: *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2010, S. 2402–2405 (siehe S. 73).

[70] Medin, D. L., Goldstone, R. L. und Gentner, D. "Respects for Similarity". In: *Psychological Review* 100.2 (1993), S. 234–278 (siehe S. 31).

[71] Mesaros, A., Heittola, T. und Palomaki, K. “Query-by-example retrieval of sound events using an integrated similarity measure of content and label”. In: *Image Analysis for Multimedia Interactive Services (WIAMIS), 2013 14th International Workshop on*. IEEE, 2013, S. 1–4 (siehe S. 73).

[72] Misdariis, N. u. a. “Environmental Sound Perception: Metadescription and Modeling Based on Independent Primary Studies”. In: *EURASIP Journal on Audio, Speech, and Music Processing* 2010 (2010), S. 1–26 (siehe S. 44).

[73] Mitrović, D., Zeppelzauer, M. und Breiteneder, C. “Features for Content-Based Audio Retrieval”. In: *Advances in Computers*. Hrsg. von Memon, A. Bd. 78. Elsevier, 2010, S. 71–150 (siehe S. 24–26, 36, 44, 45, 54, 74, 75).

[74] Moffat, D., Ronan, D. und Reiss, J. D. “An Evaluation of Audio Feature Extraction Toolboxes”. In: *Proceedings of the 18th international Conference on Digital Audio Effects (DAFx-15)*. 2015, S. 1–7 (siehe S. 35).

[75] Müller, M. *Fundamentals of Music Processing*. Cham: Springer International Publishing, 2015 (siehe S. 2, 4, 7, 8, 10, 12, 13, 21).

[76] Muzzulini, D. “Genealogie der Klangfarbe”. Diss. Universität Zürich, 2004 (siehe S. 18).

[77] Okuyucu, C., Sert, M. und Yazici, A. “Audio Feature and Classifier Analysis for Efficient Recognition of Environmental Sounds”. In: *2013 IEEE International Symposium on Multimedia (ISM)*. 2013, S. 125–132 (siehe S. 4, 7).

[78] Özcan, E., Jacobs, J. u. a. “Product sounds: Basic concepts and categories”. In: *International Journal of Design* 8.3 (2014), S. 97–111 (siehe S. 27).

[79] Pachet, F. und Aucouturier, J.-J. “Improving timbre similarity: How high is the sky”. In: *Journal of negative results in speech and audio sciences* 1.1 (2004), S. 1–13 (siehe S. 69, 75).

[80] Patil, K. und Elhilali, M. “Biomimetic spectro-temporal features for music instrument recognition in isolated notes and solo phrases”. In: *EURASIP Journal on Audio, Speech, and Music Processing* 2015.1 (2015), S. 1–13 (siehe S. 76).

[81] Patil, K. u. a. “Music in Our Ears: The Biological Bases of Musical Timbre Perception”. In: *PLoS Computational Biology* 8.11 (2012), S. 1–16 (siehe S. 76).

[82] Pedregosa, F. u. a. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), S. 2825–2830 (siehe S. 50).

[83] Peeters, G. und Deruty, E. “Sound Indexing Using Morphological Description”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 18.3 (2010), S. 675–687 (siehe S. 3, 25, 33).

[84] Peeters, G. *A large set of audio features for sound description (similarity and classification) in the CUIDADO project*. CUIDADO I.S.T. Project Report. 2004 (siehe S. 42, 43, 47).

[85] Peeters, G. u. a. “The Timbre Toolbox: extracting audio descriptors from musical signals”. In: *The Journal of the Acoustical Society of America* 130.5 (2011), S. 2902–2916 (siehe S. 25, 36, 72).

[86] Peng, Y. u. a. “Audio similarity measure by graph modeling and matching”. In: *Proceedings of the 14th annual ACM international conference on Multimedia*. ACM Press, 2006, S. 603–606 (siehe S. 33).

[87] Piczak, K. J. “ESC: Dataset for Environmental Sound Classification”. In: *Proceedings of the 23rd ACM International Conference on Multimedia*. ACM, 2015, S. 1015–1018 (siehe S. 4, 5, 54).

[88] Ricard, J. “Towards computational morphological description of sound”. Diss. Barcelona: Universitat Pompeu Fabra, 2004 (siehe S. 7, 75).

[89] Rice, S. V. und Bailey, S. M. “General-purpose real-time monitoring of machine sounds”. In: *Proceedings of the 59th Meeting of the Society for Machinery Failure Prevention Technology*. 2005, S. 545–549 (siehe S. 5).

[90] Rice, S. V., Bailey, S. M. und Corporation, C. “Searching for Sounds: A Demonstration of FindSounds.com and FindSounds”. In: *Proc. of the International Computer Music Conference*. 2004, S. 215–218 (siehe S. 5).

[91] Rocha, B., Bogaards, N. und Honingh, A. “Segmentation and timbre similarity in electronic dance music”. In: *Proceedings of the Sound and Music Computing Conference 2013*. Bd. 2. 2013, S. 754–761 (siehe S. 74).

[92] Roma, G. und Serra, X. “Querying freesound with a microphone”. In: *Proceedings of the First Web Audio Conference (Ircam, Paris, France), submission*. Bd. 39. 2015 (siehe S. 2, 34).

[93] Roma, G. u. a. “Ecological acoustics perspective for content-based retrieval of environmental sounds”. In: *EURASIP Journal on Audio, Speech, and Music Processing* (2010), S. 1–7 (siehe S. 2).

[94] Rösing, H. “Klangfarbe”. In: *Die Musik in Geschichte und Gegenwart*. Hrsg. von Finscher, L. 2. Aufl. Bd. S5. 2004, S. 138–159 (siehe S. 18).

[95] Sankiewicz, M. und Budzynski, G. “Reflections on sound timbre definitions”. In: *Archives of Acoustics* 32.3 (2007), S. 591–602 (siehe S. 18).

[96] Schaeffer, P. *Traité des objets musicaux. Essai interdisciplines*. Paris, 1966 (siehe S. 27).

[97] Schafer, R. M. *Klang und Krach. Eine Kulturgeschichte des Hörens*. Frankfurt am Main, 1988 (siehe S. 3, 73).

[98] Serrà, J. und Arcos, J. L. "An Empirical Evaluation of Similarity Measures for Time Series Classification". In: *Knowledge-Based Systems* 67 (2014), S. 305–314 (siehe S. 72).

[99] Siedenburg, K., Fujinaga, I. und McAdams, S. "A Comparison of Approaches to Timbre Descriptors in Music Information Retrieval and Music Psychology". In: *Journal of New Music Research* 45.1 (2016), S. 27–41 (siehe S. 20, 31, 75, 76).

[100] SoundCloud. *Pressemitteilung vom 04.12.2012: Next SoundCloud, Now For Everyone*. Pressemitteilung. Paris, 2012 (siehe S. 2).

[101] Spveak, C. und Favreau, E. "Soundspotter - A Prototype System for Content-Based Audio Retrieval". In: *Proceedings of the fifth international Conference on Digital Audio Effects (DAFx-02)*. Hamburg, 2002, S. 27–32 (siehe S. 4, 52).

[102] Streich, S. "Music complexity: a multi-faceted description of audio content". Diss. Barcelona: Universitat Pompeu Fabra, 2006 (siehe S. 39, 41, 42).

[103] Sunouchi, M. und Tanaka, Y. "Similarity Search of Freesound Environmental Sound Based on Their Enhanced Multiscale Fractal Dimension". In: *The Sound and Music Computing Conference 2013*. 2013, S. 715–721 (siehe S. 33).

[104] Susini, P., Houix, O. und Saint Pierre, G. "The Effect of Loudness on the Perceptual Representation of Sounds With Similar Timbre". In: *Acta Acustica united with Acustica* 101.6 (2015), S. 1174–1184 (siehe S. 28).

[105] Susini, P., Lemaitre, G. und McAdams, S. "Psychological measurement for sound description and evaluation". In: *Measurements With Persons: Theory, Methods, and Implementation Areas*. Hrsg. von Berglund, B. u. a. New York: Psychology Press, 2012, S. 227–253 (siehe S. 13, 19, 20, 28, 44).

[106] Susini, P. u. a. "Caractérisation perceptive de bruits". In: *Acoustique et Techniques* 13 (1998), S. 11–15 (siehe S. 20, 27, 53).

[107] Tan, K. *55 Great Websites To Download Free Sound Effects*. Blogeintrag. 2010 (siehe S. 2).

[108] Ueda, K. "A hierarchical structure for adjectives describing timbre". In: *The Journal of the Acoustical Society of America* 100.4 (1996), S. 2751–2751 (siehe S. 20).

[109] Valle, A. “Environmental Sound Synthesis, Processing, and Retrieval”. In: *EURASIP Journal on Audio, Speech, and Music Processing* (2010), S. 1–3 (siehe S. 7).

[110] Vickers, E. “Automatic long-term loudness and dynamics matching”. In: *Audio Engineering Society Convention 111*. Audio Engineering Society, 2001 (siehe S. 41).

[111] Virtanen, T. und Helen, M. “Probabilistic Model Based Similarity Measures for Audio Query-by-Example”. In: *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. 2007, S. 82–85 (siehe S. 4).

[112] von Bismarck, G. “Timbre of Steady Sounds: A Factorial Investigation of its Verbal Attributes”. In: *Acta Acustica united with Acustica* 30.3 (1974), S. 146–159 (siehe S. 20).

[113] Von Helmholtz. *Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik*. 13 (1968), Darmstadt. Braunschweig, 1863 (siehe S. 15, 19).

[114] Wan, C., Liu, M. und Wang, L. “Content-based sound retrieval for web application”. In: *Web Intelligence: Research and Development*. Springer, 2001, S. 389–393 (siehe S. 33, 34).

[115] Wang, A. L.-c. “An industrial-strength audio search algorithm”. In: *Proceedings of the 4 th International Conference on Music Information Retrieval*. 2003 (siehe S. 4).

[116] Ward, J. *The Student's Guide to Cognitive Neuroscience*. Psychology Press, 2010 (siehe S. 13).

[117] Weinzierl, S. “Grundlagen”. In: *Handbuch der Audiotechnik*. Hrsg. von Weinzierl, S. Springer Berlin Heidelberg, 2008, S. 1–39 (siehe S. 6, 10, 12).

[118] Wichern, G. u. a. “Segmentation, Indexing, and Retrieval for Environmental and Natural Sounds”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 18.3 (2010), S. 688–707 (siehe S. 36, 52, 73).

[119] Wold, E. u. a. “Content-based classification, search, and retrieval of audio”. In: *IEEE MultiMedia* 3.3 (1996), S. 27–36 (siehe S. 33, 52).

[120] Xue, J. u. a. “Fast query by example of environmental sounds via robust and efficient cluster-based indexing”. In: *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, S. 5–8 (siehe S. 34, 52).

[121] Zhang, T. und Kuo, C.-C. “Hierarchical classification of audio data for archiving and retrieving”. In: *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Bd. 6. 1999, S. 3001–3004 (siehe S. 4).

[122] Zhou, G.-T. u. a. “Relevance feature mapping for content-based multimedia information retrieval”. In: *Pattern Recognition* 45.4 (2012), S. 1707–1720 (siehe S. 3).

[123] Zölzer, U. *Digital Audio Signal Processing*. John Wiley & Sons Ltd, 1997 (siehe S. 42).

[124] Zölzer, U. “Signalverarbeitung, Filter und Effekte”. In: *Handbuch der Audiotechnik*. Hrsg. von Weinzierl, S. Springer Berlin Heidelberg, 2008, S. 813–848 (siehe S. 26).

[125] Zwicker, E. *Psychoakustik*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1982 (siehe S. 19).

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungen und Variablennamen

λ	Wellenlänge
$\mathcal{H}$	Hop-Size
$\mathcal{I}$	Anzahl Abtastwerte
$\mathcal{K}$	Framelänge
$\mathcal{N}$	Anzahl Frames
$\mathcal{S}$	Anzahl Klänge
$\mathcal{V}$	Anzahl Audiomerkmale
μ	Mittelwert
σ^2	Varianz
φ	Phase
A	Amplitude
c_S	Schallgeschwindigkeit
d_E	Euklidische Distanz
f	Frequenz in Hz
f_S	Abtastrate
I	Intensität in Wm^{-2}
i	Index eines Abtastwertes
$i_e(n)$	Index des Abtastwertes am Ende von Frame n
$i_s(n)$	Index des Abtastwertes am Anfang von Frame n

k	diskrete Frequenzklasse (engl. frequency bin)
L_p	Schalldruckpegel in dB
n	Frame Index
p	Schalldruck in Pa
s	Klang
T	Periodendauer
t	Zeit
v	Audiomerkmal
V_q	Merkmalsvektor von Beispielklang q
V_s	Merkmalsvektor von Klang s
v_{DynC}	Audiomerkmal DynamicComplexity
v_D	Audiomerkmal Duration
v_{ED}	Audiomerkmal EffectiveDuration
v_{LAT}	Audiomerkmal LogAttackTime
v_L	Audiomerkmal Loudness
v_{MFCC}	Audiomerkmal MFCC
v_{PC}	Audiomerkmal PitchConfidence
v_{PITCH}	Audiomerkmal Pitch
v_{SC}	Audiomerkmal SpectralCentroid
x	Audiosignal
$X(\cdot)$	Fouriertransformierte von Signal x
D1	Datenset 1
D2	Datenset 2
DCT	Diskrete Kosinustransformation

engl.	englisch
FFT	Schnelle Fouriertransformation (engl. fast Fourier transform)
FT	Fouriertransformation
KNN	k-Nearest-Neighbors-Algorithmus
MDS	Multidimensionale Skalierung
MFCC	Mel-Frequenz-Cepstrum-Koeffizienten
STFT	Kurzzeit-Fouriertransformation (engl. short-term Fourier transform)
vgl.	vergleiche