Extraktion und Beschreibung von Metadaten aus Audiodateien mittels MPEG-7

Michael Honig
Honig@ifi.lmu.de

Universität München
Amalienstrasse 17, 80333 Müchen, Deutschland

Zusammenfassung Bis heute ist es kaum möglich Nicht- Textbasierte Dateien, wie Video - oder Audiodateien, mit Hilfe einer Suchmaschine aus einer Datenbank oder dem Internet herauszufiltern. Die vorliegende Arbeit beschäftigt sich deshalb mit dem MPEG- 7 - Standard der dies in Zukunft für Video - Audiodateien möglich machen soll. Durch diesen Standard werden Dateien zunächst analysiert und daraus Metadaten, die den Inhalt dieser Dateien beschreiben, gewonnen. Die Arbeit beschäftigt sich zunächst mit den verschiedenen Möglichkeiten diese Metadaten zu erhalten. Im Anschluss daran werden die einzelnen Werkzeuge, die es ermöglichen auf die Daten zuzugreifen, vorgestellt. In diesem Zusammenhang werden einige Beispiele zur Beschreibung von Metadaten (Sprache, Sound) genannt und ihre Anwendungsbereiche beschrieben. Außerdem werden Hilfsmittel zur Extraktion und Erzeugung von Metadaten vorgestellt. Die Arbeit endet mit einem Beispiel zur Illustration einer praktischen Anwendung.

1 Einleitung

Heutzutage ist es jedem einzelnen Anwender möglich auf eine fast unübersichtliche Anzahl von Daten (z.B. mit Hilfe des Internets) zuzugreifen. Diese können dann unter Verwendung von Webbrowsern oder Suchmaschinen ohne größeren Aufwand nach den gewünschten Inhalten durchsucht und anschließend sortiert und bearbeitet werden. Eine große Ausnahme bilden jedoch Nicht - Textbasierte Dokumente (Format dieser Dateien meist in MPEG - 1, - 2, - 4). Insbesondere Audiofiles stellen ein großes Hindernis für Suchmaschinen oder Anwendungen ähnlicher Art dar. Das liegt daran, dass der Audioinhalt nicht in Textform dargestellt werden kann. somit entfällt die Möglichkeit auf den Inhalt dieser Dateien mit vorhandenen Suchmaschinen zuzugreifen. MPEG-7 versucht diese Lücke zu schließen, indem es nicht nur textbasierte Suchanfragen wie Songtexte oder Künstlernamen, sondern auch komplexe inhaltliche Anfragen (z.B.: Es sollen alle Lieder angezeigt werden die Ähnlichkeit mit einer bestimmten Melodie haben) zulässt. Wichtigste Bestandteile für diese Art von Suchanfragen sind die Metadaten. Unter Metadaten versteht man Daten die den Inhalt oder den Aufbau einer Audio - oder Videodatei beschreiben. Diese Metadaten werden mit Hilfe von Low -Level Deskriptoren für elementare Audiomerkmale (Lautstärke, Tonhöhe usw.) und mit High - Level Deskriptoren für abstraktere Audiomerkmale (Melodien, Tonfolgen usw.) aus der Audiodatei extrahiert.

Die Arbeit ist folgendermaßen gegliedert: Das nächste Kapitel gibt einen allgemeinen Überblick über die Extraktion von Audiodaten. Im dritten Kapitel werden die vorhandenen Audiostrukturen vorgestellt. Anschließend werden einige Anwendungsbeispiele genannt. Den Abschluss bildet ein Fazit.

2 Überblick über die Extraktion von Audiodaten

Mit MPEG-7 wird ein Standard präsentiert der es ermöglicht, Metadaten mit den Inhalten von Mediadaten zu assoziieren. Im Gegensatz zu früheren MPEG-Standards wird hier kein En- bzw. Dekodieren standardisiert oder näher spezifiziert. Der MPEG-7 Standard besteht aus so genannten Deskriptoren und Descritpion Schemes.[1] Deskriptoren sind Umschreibungen von Metadaten die mit einem einzelnen oder mehreren Zeitintervallen oder sogar dem ganzen Signal einer Audiodatei in Verbindung gebracht werden. Ein Description Scheme ist ein Verbund von Deskriptoren und beschreibt ihren Zusammenhang untereinander. Diese Schemes basieren auf XML , wie im folgenden Codebeispiel zu sehen ist

```
<AudioDescriptionScheme xsi:type="PercussiveInstrumentTimbreType"
  <LogAttackTime>
    <Scalar>-1.683017</Scalar>
  </LogAttackTime>
  <SpectralCentroid>
    <Scalar>1217.341518</Scalar>
  </SpectralCentroid>
  <TemporalCentroid>
    <Scalar>0.081574</Scalar>
  </TemporalCentroid>
</AudioDescriptionScheme>
```

Abbildung 1. Beispiel für ein AudioDesription Scheme
[2]

Der MPEG-7 - Audio - Standard stellt Description Schemes als fundamentale Konstrukte für anwendungsbasierte Werkzeuge zur Verfügung, um mit Inhalten von Audiodaten zu arbeiten Da aber die Möglichkeit besteht, daß die vorgegebenen Werkzeuge nicht für jede Anwendung passend sind, existiert eine so genannte Descritpion Definition Language (DDL) [3] die ebenfalls auf XML basiert. Dadurch ist es möglich neue und auf eine Anwendung zugeschnittene Descritpion Schemes zu erstellen. Wie eine solche Extraktion von Audiodateien durchgeführt wird, wird im Folgenden genauer erklärt.

Da eine Audiodatei meist aus mehreren verschieden akustischen Inhalten besteht, wie Geräusche, Töne oder Sprache, gibt es verschiedene Möglichkeiten nach eben diesen Merkmalen mit den von MPEG-7 zur Verfügung stehenden Werkzeugen zu suchen und diese dann für eine Analyse oder für weitere Anwendungen zu extrahieren. Man kann die Inhaltsextraktion grob in folgende vier Kategorien unterteilen: [4]

Suche mit Hilfe von Metadaten (keyword-based search): Bei der ersten Kategorie wird der eigentlich Inhalt des Audiosignals nicht berücksichtigt und nur mit Hilfe von textueller Zusatzinformation gesucht. So ist jede Audiodatei mit einem oder mehreren Attributen versehen und auf diese Attribute kann dann mit klassischen Textretrievalmethoden zugegriffen werden. Beispiele für solche Attribute sind: Bandname, Entstehungsdatum, Liedname Genre usw.

Suche mit Hilfe von akustischen Merkmalen (content-based similarity search): Die zweite Kategorie beschreibt die Signalinformationen mit Hilfe von Merkmalen die den Audioinhalt eindeutig beschreiben. Beispiele für solche Merkmale sind Lautstärke und Tonhöhe. Diese werden mit Hilfe des Audio Description Frameworks (Kapitel 3) direkt aus den Audiodaten extrahiert.

Suche mit Noten bzw. Ton Intervallen (search by humming): Für die dritte Kategorie wird eine Ähnlichkeit hinsichtlich des Taktes, des Tempos oder der Noten definiert. Bei der Notenerkennung ergibt sich bei Musikstücken mit mehreren Instrumenten jedoch eine nicht - triviale Problematik, die gelöst werden muss: Die verschiedenen Instrumentstimmen überlagern sich zum Teil gegenseitig bzw. werden zusätzlich von Gesang überlagert. Aus diesem Grund speichern die meisten Ansätze Daten in Formaten, die zum Beispiel darin vorkommende Instrumente oder Noten beinhalten (MIDI). Um das Problem der Melodieerkennung zu lösen werden die Noten nicht direkt verglichen sondern es werden die Notenintervalle gespeichert mit der zusätzlichen Information, ob der nachfolgende Ton höher oder tiefer ist, also eine Umwandlung der Notenfolge in eine Differenzfolge. Diese Zuweisung erfolgt für jede Note einzeln und auch nur im Verhältnis zur vorhergehenden Note. Ein D wird gespeichert wenn die vorherige Note höher war, ein U wenn die vorhergehende Note tiefer war und ein S wenn die Notegleich sind. Somit kann eine Anfrage mittels Summen der Melodie oder durch die Eingabe von Zeichenfolgen erfolgen. Ein Anwendungsbeispiel dazu wird in Kapitel 4.1 beschrieben.

Suche in gesprochenem Text (speech regonition): In der vierten Kategorie wird mit Hilfe von Textretrieval Methoden versucht Phoneme oder ganze Wörter aus einem gesprochenen Text zu extrahieren und zu verarbeiten.

Dies erfolgt mit Hilfe von High- Level Audio Deskriptoren und Spracherkennungssystemen. Um diese besser zu veranschaulichen wird in Kapitel 3 auf die Spracherkennung detailliert eingegangen.

2.1 MPEG-7 Audio Strukturen

Um den auditiven Inhalt von Audiodateien zu beschreiben stellt MPEG-7 Audio generell zwei Arten von Strukturen bereit. Einerseits das Audio Description Framework das mit den Low - Level Deskriptoren elementare auditive Merkmale beschreibt und so genannte High - Level Deskriptoren die auf diesen extrahierten Daten aufbauen und mehr für eine anwendungsspezifische Extraktion konzipiert sind.

2.2 Low - Level Deskriptoren

Grundsätzlich werden Low -Level auditive Merkmale mit Hilfe von Deskriptoren auf zwei Arten beschrieben. Entweder man teilt z. B. ein Musikstück in gleichlange Zeitintervalle ein und hält aus jedem Intervall ein repräsentativen Wert fest, oder man teilt das Musikstück in Segmente ein und beschreibt die Ähnlichkeiten bzw. Differenzen der Segmente. Insgesamt sind in MPEG-7 Audio siebzehn Audio - Deskriptoren definiert.Dargestellt in Abbildung 3. Darunter sind zeitspezifische- und spektrale Deskriptoren, also Deskriptoren die die Wellenform der Signale betrachtet. Diese Deskriptoren sind in folgende Gruppen eingeteilt:

Basis - Deskriptoren: Die Audio Basis - Deskriptoren beinhalten nur Wellenformen (Skalare), die einen bestimmten Zeitabschnitt eines Signals beschreiben (z. B. minimale und maximale Tonfrequenz). Siehe Figur 2.

Basisspektrum - Deskriptoren: Die vier Basisspektrum - Deskriptoren werden für die Beschreibung verschiedener Frequenzspektren verwendet und können durch einfache Zeitfrequenzanalysen erzeugt werden.

Signalparameter - Deskriptoren: Die beiden Signalparameter - Deskriptoren beziehen sich hauptsächlich auf periodische und quasi-periodische Signale. Damit kann z. B. die Harmonie eines Signals beschrieben werden.

Temporäre Klangfarbe - Deskriptoren: Beide Temporäre - Klangfarben - Deskriptoren beschreiben zeitliche Charakteristiken von Tonsegmenten und sind besonders für die Beschreibung der Tonqualität unabhängig von Tonhöhe und Lautstärke geeignet.

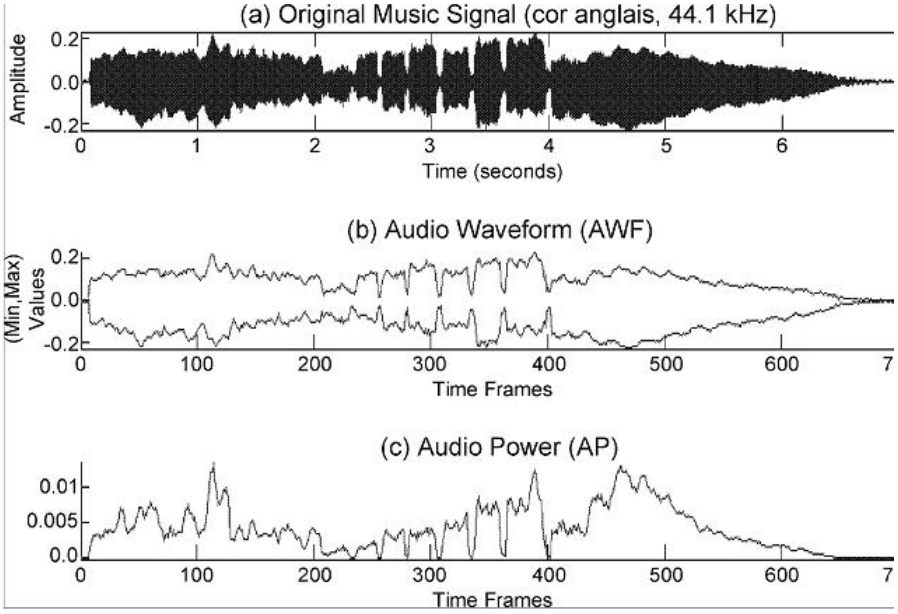

Abbildung 2. Mpeg-7 Basis - Deskriptoren Signal a zeigt das ursprüngliche Signal an, b zeigt die Wellenform an (nach Behandlung mit einem Basisdeskriptor), c die Lautstärke des Signals (ebenfalls nach Behandlung mit einem Basisdeskriptor)
[2]

Spektralklangfarbe - Deskriptoren: In der Gruppe Spektralklangfarbe gibt es fünf Deskriptoren, die Spektralmerkmale in einem linearen Frequenzraum beschreiben. Hier geht es besonders um die Wahrnehmung der Harmonie eines auditiven Signals.

Spektralbasis- Deskriptoren: Die beiden letzten Low - Level Audio - Deskriptoren gehören zur Gruppe Spektralbasis und repräsentieren Projektionen aus einem mehrdimensionalen Spektralraum in einen Spektralraum geringerer Dimension, um Kompaktheit und Wiedererkennung von auditiven Inhalten zu unterstützen.

Es gibt darüber hinaus noch ein spezielles Audio Segment, mit dem Namen Silence Segment (Ruhesegment) Dieses Segment beinhaltet die Semantik "Ruhe", es zeigt also an, wann in der Datei keinerlei Geräusche vorkommen, also "Ruhe" herrscht. Der Anwendungsbereich dieses Segments Liegt zum Beispiel in der Trennung von Dateien, da es von Vorteil ist, wenn sich keine Geräuschkulisse am Trennungspunkt befindet und einen unmerklichen Übergang zu ermöglichen.

2.3 High-level Deskriptoren und Deskription Schemata

Die "Low - Level" Audio - Deskriptoren beschreiben auditive Merkmale auf einer sehr geringen Abstraktionsebene. Um auditive Daten mit einer höheren

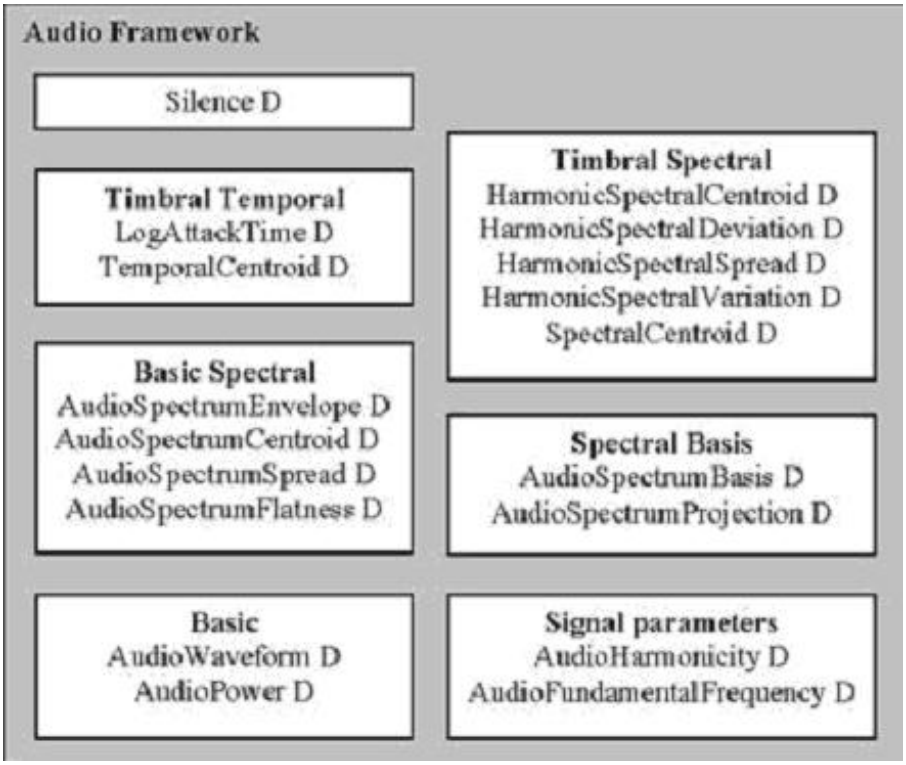

Abbildung 3. Zusammenfassung des Audio Frameworks [5]

Abstraktion zu beschreiben, wurden in MPEG-7 Audio die "High - Level" Audio - Deskriptoren und Deskriptions Schemata definiert. Darin enthalten befindet sich das Audio Signatur description scheme und die description schemes für Klangfarben und Melodie, die für die Unterstützung von Abfragen durch den Menschen geeignet sind. Weiterhin sind allgemeine Tonerkennungs- und Tonerkennungsindizierungswerkzeuge vorhanden, die zur Indizierung und Kategorisierung von Tönen dienen. Hierfür werden Wahrscheinlichkeitsmodelle - das Hidden Markov bzw. Gaussian Mixture Modell - zur Klassifizierung von Soundeffekten verwendet. Beide Modelle werden im folgenden Kapitel genauer erklärt. Die Beschreibungswerkzeuge für gesprochenen Inhalt erlauben detaillierte Beschreibungen von gesprochenen Wörtern innerhalb eines auditiven Datenstromes. Diese Werkzeuge sind auch für eine enge Zusammenarbeit mit den Automatic Speech Recognition (ASR) Werkzeugen geeignet. Leider weisen ASR Werkzeuge heutzutage noch gewisse Lücken in der Spracherkennung bezogen auf Wortschatz und Dialekt auf, somit bleibt abzuwarten, inwieweit MPEG-7 zusammen mit ASR Werkzeugen in Zukunft genutzt wird. Die genaue Arbeitsweise von ASR wird in Kapitel 3 genauer beschrieben.

3 Spoken Content:

Eine der intuitivsten und offensichtlichsten Arten von Metadaten die aus einem Multimediadokument extrahiert werden können ist die Sprache (Spoken Content).Sie kann sowohl als Schlüsselwortsuche oder auch als Input für eine weitere Metadatensuche oder als Identifikator für inhaltliche Themen verwendet werden. Ein Beispielszenario in dem Sprache als Metadatum verwendet wird: Es wäre möglich Fotos mit verbalen Kommentaren oder Bemerkungen zu versehen, die dann für eine spätere, einfachere Wiedererkennung des Inhaltes oder der gerade gedachten Assoziationen von großem Nutzen sind. Dieses Foto wird auf einem Server gespeichert und die Bemerkungen werden mit Hilfe von automatic speech recognition (ASR) dekodiert und als Spoken Content Metadaten gespeichert. Siehe Abbildung 4.

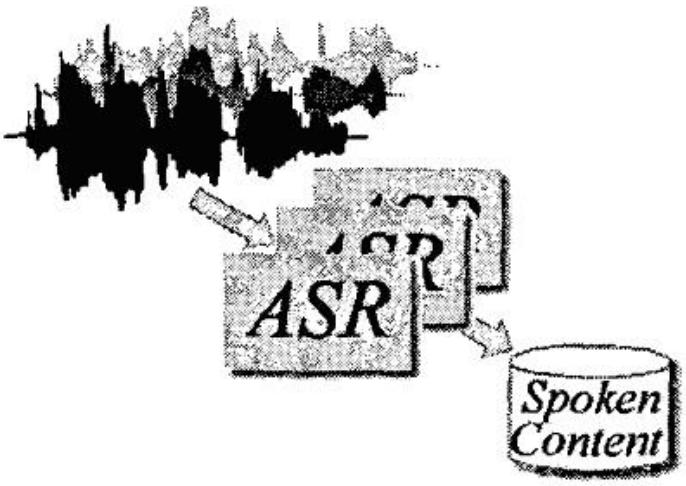

Abbildung 4. Sprache (Spoken Content) wird mit Hilfe von ASR aus einem oder mehreren Audiostreams gefiltert.

[6]

Somit besteht nun die Möglichkeit auch nach verbalen Inhalten die Fotos zu durchsuchen oder zu ordnen. Bei diesem Anwendungsfall kommen die Nachteile von ASR noch nicht sehr zum tragen. Will man jedoch gesprochene Daten zum Beispiel aus Filmen mit mehreren Sprechern oder sogar verschiedenen Sprachen extrahieren, stoßen gewöhnliche ASR Systeme schnell an ihre Grenzen. ASRs speichern die extrahierten Daten in form einer Graphenstruktur ab. Siehe Abbildung 5.

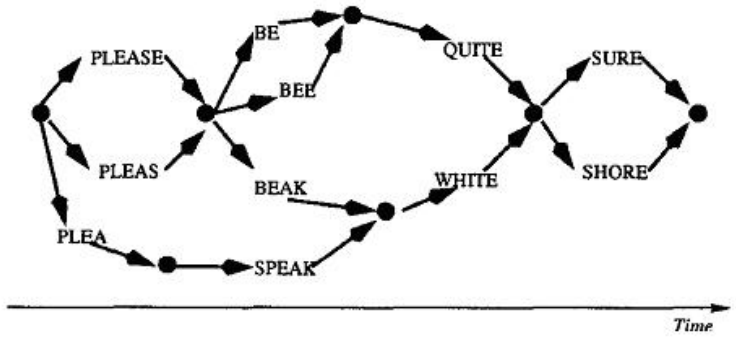

Abbildung 5. Hypothetische Graphenstruktur die den Satz please be quite sure darstellt. [6]

So kann eine relativ große Anzahl von Möglichkeiten in kompakter Form dargestellt werden. Bei sehr großen Datenmengen wird dies allerdings zu einem Problem. Auch Verfügen ASRs in der Regel über einen Wortschatz von 20. - 60.000 Worten. Dies bedeutet allerdings das viele Eigen- Personen oder Städtenamen nicht enthalten sind. Alternativ gäbe es die Möglichkeit die Darstellung nicht wortbasiert sondern auf Silben oder lautsprachlichen Ausdrücken zu basieren. Jedoch führt dieses bei mehreren Sprachen auch wieder zu Überschneidungen. Um diesem Entgegenzuwirken verwendet MPEG-7 den SpokenContent Descriptor, eine Mischung aus einer Wort und Laut(phonetischen) Darstellung, wie in Abbildung 6 dargestellt.

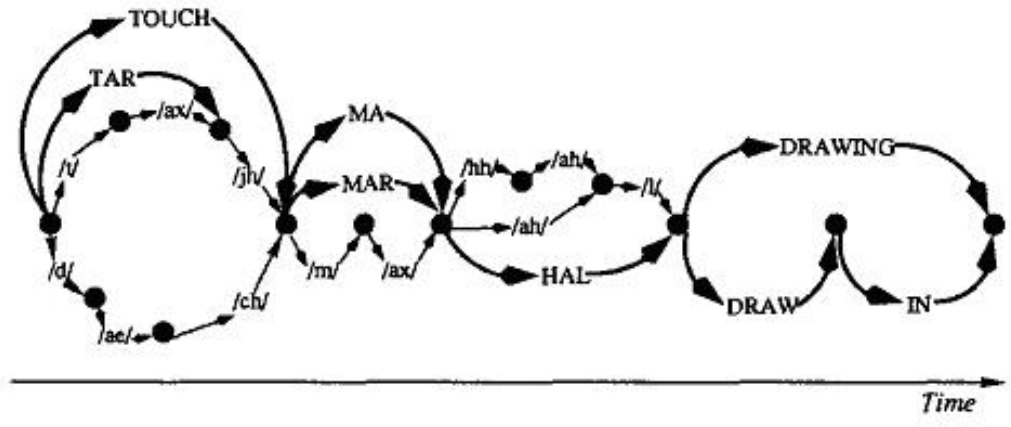

Abbildung 6. Hypothetische Graphenstruktur die Wort und Lautliche Darstellung kombiniert. Der Satz lautet Taj Mahal drawing wobei der Begriff Taj Mahal unbekannt ist [6]

Dieser wird jedoch mit zusätzlichen Metadaten versehen, die dann im dem Header gespeichert werden.

Jedem Sprecher (Speaker) wird eine Sprache, ein Wort und ein Silben/Lautsprachenlexikon zugeordnet. Phone Statistics (siehe Abbildung 7) hilft durch die Wahrscheinlichkeitsverteilungen der einzelnen Laute (ca 45 im Englischen) die Auswahlmöglichkeiten weiter einzuschränken. Somit wird durch die Verwendung von zusätzlichen Metadaten und nicht nur durch die Verwendung von reinen Low - Level Deskriptoren sichergestellt, dass die gewünschte Information extrahiert werden kann.

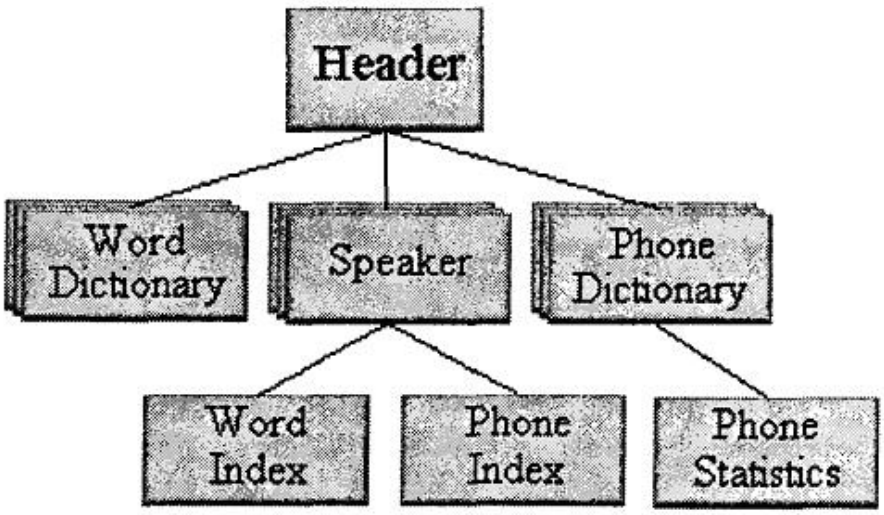

Abbildung 7. Schematische Darstellung des Header In der MPEG-7 Struktur [6]

4 Sound Classification

Durch die immer größer werdende Zahl von digitalen Audiodaten auf die ein Nutzer zugreifen kann, entstand die Notwendigkeit diese Daten zu Klassifizieren und zu sortieren. Die für die Klassifikation benötigte Information sollte nicht durch die Qualität der Daten oder Ihrer Formate eingeschränkt werden. Da allerdings eine Beschränkung auf reine Low - Level Deskriptoren nicht zum gewünschten Ergebnis führen würde, da die entstehende Datenmenge einfach zu groß wäre, werden zusätzlich die High - Level Tools SoundModel und SoundClassificationModel und das Hidden Markov Model (HMM)[7] und der Viterbi - Algorithmus verwendet[8]. Allgemein läuft eine Klassifikation nach folgendem Schema ab: Zunächst muss mit Trainingsdaten eine Anzahl von SoundModel - Deskriptoren extrahiert werden, mit denen dann der SoundClassifcationModel-Deskriptor erstellt wird. Dieser besteht aus mehreren

verschiedenen Geräuschklassen die untereinander in Verbindung stehen und in einer Baumstruktur dargestellt werden. Dies ermöglicht eine semantische Verbindung zwischen den einzelnen Kategorien.Siehe Abbildung 8.

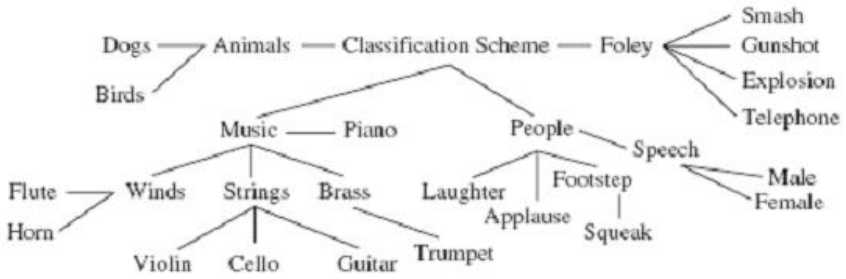

Abbildung 8. Beispiel für ein SoundClassificationModel description scheme [9]

Anschließend wird ein vorhandener AudioSpectrumEnvelope-Deskriptor mit dem Hilfe eines AudioSpectrumBasis-Deskriptor auf einen reduzierten Merkmalsraum abgebildet. Diese Abbildung erfolgt mit dem AudioSpectrumEnvelope-Deskriptor der jedes Audiosegment in kurze Zeitintervalle unterteilt und anschließend die dort enthaltenen Welleformen auf n-dimensionale Vektoren abgebildet. Selbst bei einer sehr niedrigen Sampelrate würde eine Klassifikation immer noch einen enormen Rechenaufwand darstellen. Um dieses Problem zu umgehen wird der ursprüngliche Vektor mittels einer Singulärwertzerlegung (SVD)[10] auf eine kleinere Basis reduziert, wie in Abbildung 9 gezeigt.

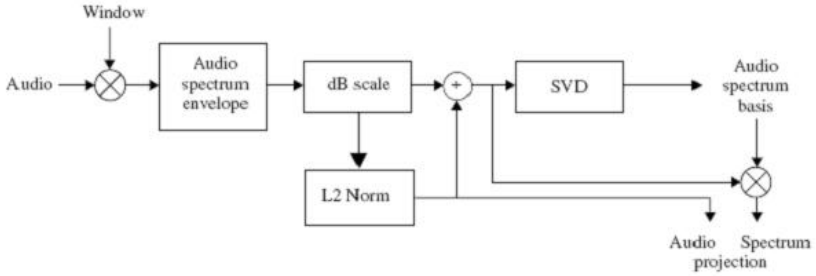

Abbildung 9. AudioSpectrumBasis und AudioSprektrumProjection Descriptor [?]film)

Dann wird mit Hilfe des Hidden-Markov Modells (HMM) und dem Viterbi-Algorithmus der beste Pfad und seine Wahrscheinlichkeit berechnet. Dadurch daß sich das Spektrum eines Audiosignals mit der Zeit verändert und diese Veränderungen für das Signal typisch ist (so erkennt man bei einem

bellenden Geräusch das es sich um einen Hund handelt und dieses Geräusche eine sehr große Ähnlichkeit untereinander aufweisen), kann diese Eigenschaft zur Klassifizierung verwendet werden. Das HMM wird auf die Klassen des SoundClassificationModel Deskriptor angewendet, da jede Klasse verschiedene Zustände hat und jeder dieser Zustände durch Wahrscheinlichkeitsverteilungen definiert wird. Ein HMM zeichnet sich dadurch aus das es aus einer Kette von Zuständen und Übergangswahrscheinlichkeiten besteht. Die Zustände sind von außen nicht zu erkennen (Hidden) und ein Folgezustand hängt nur von seinem vorherigen Zustand ab.Abbildung 10.

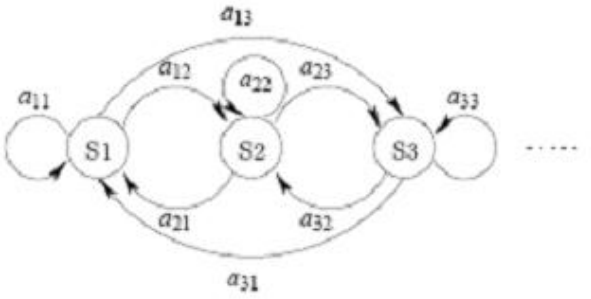

Abbildung 10. Beispiel für eine Markov - Kette
[11]

Bei der Klassifikation das Klangbild durch die wahrscheinlichste Abfolge von Zuständen einer Klasse beschrieben. Nach Anwendung des HMM hat jede Klasse nun eine Anfangswahrscheinlichkeit, mehrere Zustände und eine Zustandsübergangsmatrix.. Das ermöglicht daß für jeden Vektor für jedem Zustand eine Wahrscheinlichkeit berechnet werden kann. Der Viterbi - Algorithmus dient nun dazu für die wahrscheinlichste Folge von Zuständen zu errechnen. Als Ergebnis liefert er dann eine Vereinfachte Markov - Kette, siehe Abbildung 11.

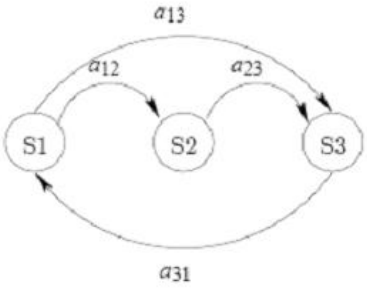

Abbildung 11. vereinfachte Markov - Kette
[11]

In einem letzten Schritt wird dann das SoundModel das am besten passt, also das mit der höchsten Wahrscheinlichkeit, ausgewählt und dem Klangbeispiel dieser Klasse zugeordnet, was mit Abbildung 12 verdeutlicht wird.

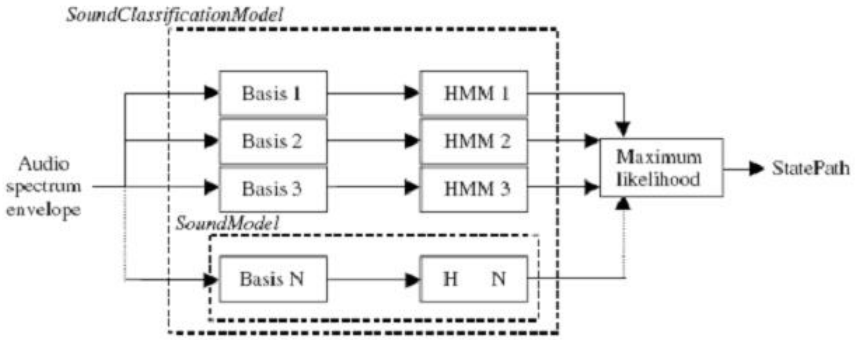

Abbildung 12. Aufbau einer Klassifizierung
[6]

5 Beispiele

Query by humming Ein typisches Beispiel für Query by Humming besteht aus folgendem Aufbau: Es existieren ein oder mehrere Media - Server. Auf diesen Servern liegen mit MPEG-4 komprimierte und kodierte Audiodateien. Weiterhin gibt es einen weitern Server der mit dem erstgenannten verbunden ist und auf dem eine Datenbank mit MPEG-7 Metadaten abgelegt ist. In dieser Datenbank ist die Melodie der Lieder, der auf dem MPEG-4 Server abgelegten Musikstücke abgebildet, und sie enthält auch eine eindeutige Zuweisung der einzelnen Metadaten zu den korrespondierenden Musikstücken. Der Benutzer startet nun eine Anfrage und summt eine Melodie bzw. gibt die Zeichenfolge an (siehe Abbildung 13). Dies bewirkt eine Anfrage bei dem MPEG-7 Server und das abgetastete Audiosignal wird nun an diesen weitergeleitet. Dort werden die extrahierten Metadaten mit den vorhandenen Einträgen in der Datenbank verglichen und, bei Übereinstimmung oder vorher spezifizierter Ähnlichkeit, werden die gefundenen Daten der Lieder oder auch weitere Informationen wie Künstlername oder Albumtitel an den Anfragesteller zurückgeschickt. Dieser kann sich nun die auf dem MPEG-4 Server gespeicherten Lieder, die zu seiner Anfrage passen, herunterladen oder abspielen.

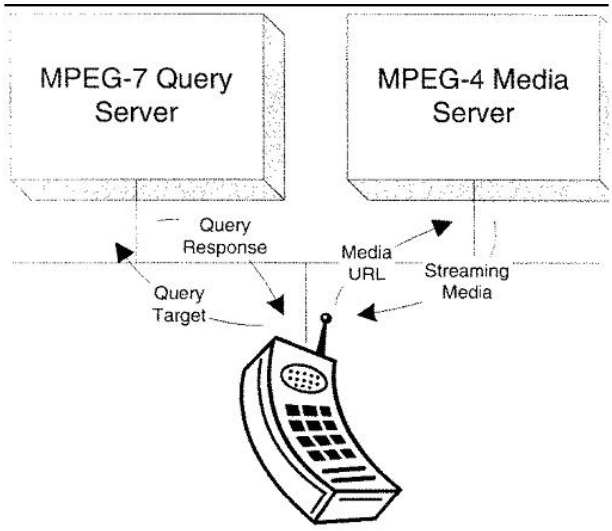

Abbildung 13. Mpeg-7 client-server Darstellung [12]

Assisted Consumer-Level Audio Editing Ein weiterer denkbarer Anwendungsbereich wäre die Möglichkeit das die extrahierten Metadaten einer Audiodatei für die Erstellung eines intuitiven Musikeditors verwendet werden können. So könnten die Deskriptoren die Darstellungen der einzelnen Low - Level Audioeigenschaften wie Spektralform für den Benutzer leichter Verständlich machen oder so darstellen das der Benutzer sehr leicht damit arbeiten kann oder mit diesem Editor Audiodatei bearbeiten kann, wie zum Beispiel bei verschiedenen Liedern die Lautstärke aneinander anzupassen.

6 Zusammenfassung

Abschließend ist zu sagen, das MPEG-7 mit den in dieser Arbeit vorgestellten Werkzeugen es möglich macht Audiodaten mit Metadaten zu indizieren und somit ein großes Spektrum für mögliche Anwendungen wie zum Beispiel die Suche nach ähnlichen Musikstücken (AudioID)[13] oder die Suche nach gesprochenen Nachrichten im Radio oder Fernsehen, die als Ergebnis eine Audio - oder Videodatei liefert, zur Verfügung stellt. Gegenwärtig gibt es jedoch kaum Anwendungen die diesen Standard unterstützen und es muss noch einige Grundlagenforschung betrieben werden um die bestehenden Problematiken wie mangelnde Konformität der Semantik zu beheben und somit die vorhandenen Nachfragen von Nutzern und Unternehmen nach solchen Anwendungen zu erfüllen.

Literatur

1. MPEG. (http://www.chiariglione.org/mpeg/)
2. H.-G. Kim, N.M., Sikora, T.: MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval. (2005)
3. Adam T. Lindsey, Ian Burnett, S.Q., Jackson, M.: Fundamentals of Audio Descriptions. (2002)
4. Härder, P.D.: Audio / Video retrieval. (2002)
5. Martínez, J.M.: MPEG-7 Overview (version 10). (2004)
6. JPA Charlesworth, P.G.: Spoken content metadata and MPEG-7. (2000)
7. Wikipedia: Hidden markov model. http://de.wikipedia.org/wiki/HMM (2006)
8. Wikipedia: Viterbi - algorithmus. http://de.wikipedia.org/wiki/Viterbi-Algorithmus (2005)
9. michael a.casey: sound classification and similarity. (http://de.xenia.media.mit.edu/ mkc/caseyChapter.pdf)
10. Wikipedia: Singulärwertzerlegung (1995)
11. Schönwandt, T.: (Filmszenenklassifizierung anhand der Tonspurmit Hilfe des MPEG-7 Standards)
12. S Quackenbush, A.L.: Overview of MPEG-7 audio. (2001)
13. FrauenhoferIDMT: (Audio-identifikation (audioid))