WIEDERERKENNUNG VON AUDIOSIGNALEN - VERFAHREN UND ANWENDUNGEN

Christian-M. Westendorf
GeSiM mbH Dresden
gesim@t-online.de

Abstract. Der vorliegende Beitrag beschäftigt sich mit Verfahren und Anwendungsmöglichkeiten der Wiedererkennung von Audiosignalen. Charakteristisch ist dabei, dass dabei nur geringe Abweichungen zwischen dem zu identifizierenden Muster und den Referenzdaten zugelassen werden. Dadurch gestaltet sich die eigentliche Erkennung relativ einfach. Andererseits werden aber hohe Anforderungen an die Effizienz und Zuverlässigkeit dieser Verfahren gestellt, die entsprechende Signalverarbeitungsmethoden benötigen. So erfordert ein Teil der Anwendungen Echtzeit-Lösungen, andere Anwendungen sind wiederum durch sehr große Datenbasen charakterisiert.

Verfahren der beschriebenen Art kommen bereits in vielfältigen Applikationen zum Einsatz, insbesondere bei der Überwachung von Copyrights und Werbeverträgen, zur Identifikation von Musiktiteln, in der Marktforschung oder der Ausblendung bestimmter Abschnitte in Rundfunk- und Fernsehprogrammen.

Neben einer kurzen Betrachtung grundlegender Ansätze wird ein im Einsatz befindliches System in seinen Hauptkomponenten vorgestellt. Dieses System ermöglicht es, vom Anwender vorgegebene Soundmuster begrenzter Länge - z.B. Musik, Sprache, Geräusche, Filmausschnitte, Werbeeinblendungen, Jingles etc. im laufenden Audiosignal von Rundfunk- und Fernsehprogrammen in Echtzeit zu identifizieren. Die Lösung besteht aus einer Programmkomponente, die die Generierung von spezifischen Datenbanken ermöglicht, in denen die Referenzdaten in komprimierter Form gespeichert werden, sowie einer Detektorkomponente, die in Echtzeit läuft und die von einer Videokarte gelieferten Soundsignale verarbeitet.

1 Einführung

Es gibt neben der Aufgabe, stark variierende Signale - wie sie insbesondere die Sprache darstellt - zu erkennen, auch zahlreiche Anwendungen, in denen eine minimale oder zumindest keine Quellenvarianz der zu erkennenden Audiosignale gegeben ist. Solche Signale entstehen immer dann, wenn die Quelle eine "Konserve" ist. Praktisch relevante Beispiele sind das Rundfunk- und Fernsehprogramm oder Musikkonserven jeder Art. Im Falle von Rundfunk- und Fernsehsendern kann je nach Ausrichtung des Senders fast die gesamte Sendezeit mit dem wiederholten Abspielen vorproduzierter Audiokonserven gefüllt sein. Spiel- und andere Filme sind ebenfalls dazu zu zählen, soweit sie über einen wiedererkennbaren Audioanteil verfügen. Die hier interessierenden Rundfunk- und Fernsehsender lassen sich darüber hinaus mit quasi gleichbleibend guter Qualität empfangen (Kabel, Satellit, Digitales Fernsehen).

Im Falle individuell abgespielter Tonträger, bei einer Übertragung über Handys oder bei schlecht empfangbaren Rundfunksendern kann allerdings eine erhebliche akustische Störung bzw. eine Varianz des Übertragungskanals hinzukommen.

2 Anwendungshintergrund

An erster Stelle der Anwendungen ist die *Überwachung von Rechten* an ausgestrahlten Werken zu nennen. Der Rechteinhaber hat ein Interesse daran, über die Aussendung seiner Werke möglichst komplett informiert zu sein, um die entsprechenden Vergütungen einfordern zu können. Die damit ins Spiel kommenden rechtlichen Aspekte werfen natürlich die Frage nach der Beweiskraft der Erkennungsergebnisse auf. Weitestgehende Rechtssicherheit kann hier nur durch einen kompletten Mitschnitt des Programms durch entsprechend autorisierte Einrichtungen gewährleistet werden.

Auf der anderen Seite haben z.B. Auftraggeber von Werbeausstrahlungen auch ein Interesse daran, dass der Werbespot zur vereinbarten Zeit, in voller Länge und mit der in Rechnung gestellten Häufigkeit ausgestrahlt wird.

Neben der urheberrechtlichen Motivation können aber auch *medienstatistische Erhebungen* das Ziel der Analyse sein. Interessant ist in diesem Zusammenhang das Schweizer System "Radiocontrol", bei dem die Soundaufnahme und -kompression in einer Armbanduhr untergebracht ist. Die nicht rückkodierbaren Aufnahmen werden dann offline ausgewertet und mit allen interessierenden Radio- und TV-Programmen verglichen.

Sofern es lediglich um die Auffindung eines begrenzten Vorrats von Audiomustern geht, kann die vorzuhaltende Datenbank vergleichsweise klein sein. Dafür ist das Audiosignal selbst unbegrenzt und meist online zu verarbeiten.

Eine weitere Gruppe von Anwendungen bezieht sich auf die *Suche in z.T. sehr großen Audiodatenbanken*. Hierbei hat der Nutzer ein Interesse daran, zu einem meist fragmentarischen Audiosignal genaue Angaben über die Herkunft bzw. über Titel, Autor, Verlag etc. zu erhalten. Oder aber es soll eine größere Menge digital vorliegender Audiosignale nachträglich klassifiziert werden. Diese Art der Suchaufgabe ist durch ein kurzes und als Ausschnitt des Originals vorliegendes Audiosignal sowie sehr große Audiodatenbanken gekennzeichnet. Insofern können solche Aufgaben nur in Zusammenarbeit mit Einrichtungen bewältigt werden, die über solche Datenbanken bzw. über das Rohmaterial und die Kapazität zur Erstellung solcher Datenbanken verfügen.

3 Verfahren

Grundlegend unterscheidet man zwischen *Fingerprinting* und *Watermarking*, wenngleich hier die Begriffsbildung nicht immer einheitlich ist [11]. Fingerprinting-Verfahren sind dadurch gekennzeichnet, dass die zur Wiedererkennung verwendete Information aus dem Audiosignal extrahiert wird, ohne dieses in gezielter Weise zu verändern. Bei Watermarking-Verfahren werden (auditiv nicht wahrnehmbare) Anteile zum Signal hinzugefügt, anhand derer die Erkennung durchgeführt wird.

Fingerprinting-Systeme haben den Vorteil, dass sich Fingerprints auch von bereits in Gebrauch befindlichen Audiomustern herstellen lassen, ohne dass eine Übereinkunft mit dem Produzenten oder Verteiler erforderlich ist. Um Watermarking-Technologien in den genannten Applikationen in großem Umfang einsetzen zu können, ist eine entsprechende Standardisierung unerlässlich.

Auch bei Fingerprinting-Verfahren wird eine Standardisierung im Rahmen des MPEG-7-Standards angestrebt. Die im Markt operierenden Firmen verwenden in der Regel jedoch proprietäre Formate zur Darstellung der Referenzmuster.

Die Erstellung der sog. Fingerprints ist auf verschiedenen Ebenen möglich. Die Skala reicht von der Beschreibung des allgemeinen Charakters von Musikstücken (z.B. zur Suche nach ähnlich klingenden Stücken) über die Erkennung von Musikstücken (unabhängig von deren konkreter Realisierung) bis hin zum exakten Match, d.h. der Erkennung genau einer Realisierung, die letztlich von derselben Quelle (Tonaufnahme) stammt und bestenfalls mehrere Kopier- und Übertragungsschritte durchlaufen hat. Bei der vorgesehenen Applikation ging es jedoch nicht vordergründig um Musiktitel, sondern um die Wiedererkennung von Soundmustern, wie sie im Rundfunk- und Fernsehbereich typischerweise auftreten (z.B. Werbespots oder Filme mit einem Gemisch aus Musik, Sprache und Geräusch).

Auch die Forderung, dass das Soundmuster anhand eines kurzen Ausschnittes zu erkennen ist (z.B. bei der Identifikation von Musiktiteln), ist hier nicht zu stellen. Vielmehr geht es um die Verifikation einer weitgehend vollständigen Präsentation des Audiomusters.

Prinzipiell beruhen Verfahren zur Wiedererkennung von Audiomustern auf einem Vergleich von Sequenzen, die das Audiosignal in geeigneter Weise repräsentieren. Jedes Element der Sequenz bildet einen bestimmten Abschnitt (Frame) des Signals so ab, dass ein Mustervergleich mit einem Frame in der Datenbank effektiv durchgeführt werden kann. Im Gegensatz zur Spracherkennung kommt es aber nicht unbedingt darauf an, bestimmte Signaleigenschaften im Interesse der abstrahierenden Erkennung zu unterdrücken, sondern unterscheidende Charakteristika sollen weitgehend erhalten bleiben.

Bei robusten Verfahren sind allerdings eine Reihe möglicher Signalveränderungen zu kompensieren, insbesondere Lautstärkeveränderungen, Audiokodierungen, Veränderungen des Frequenzgangs, Bandbegrenzung, Dynamikkompression oder zeitliche Dehnungen/Kompressionen. Wenn die Aufnahme mit einem Mikrofon erfolgt, kommt der Einfluss der Mikrofon-Lautsprecherkopplung hinzu (z.B. Echoeffekte). Die vorliegende Applikation geht jedoch davon aus, dass die Wiedergabe in stets gleichbleibender Qualität erfolgt und dass die Referenzmuster und die zu identifizierenden Muster denselben Übertragungskanal durchlaufen.

4 Das implementierte Verfahren

4.1 Die Erkennung

Das implementierte Verfahren besteht aus folgenden Schritten:

- Signalabtastung, Digitalisierung
- FFT-Signalanalyse mit konstanter Framebreite, gefenstert und bandbegrenzt
- Lineare Merkmaltransformation und -reduktion auf 6 Merkmale
- Vektorquantisierung mittels eines zuvor bestimmten Codebuchs (nur Referenzmuster)
- DP-ähnliches Matching zwischen den Mustersequenzen und dem Eingangssignal
- Musterspezifische Schwellwertentscheidung über das Vorliegen eines Matches, Hypothesenbildung
- Nachverarbeitung der Hypothesensequenzen, Entscheidung über die Ausgabe

Das Audiosignal selbst wird mittels einer handelsüblichen TV-Karte gewonnen, die auf den gewünschten Sender abzustimmen ist. Die Digitalisierung des Signals erfolgte mit den in jedem PC verfügbaren Soundblaster-kompatiblen Audioschnittstelle (Abtastfrequenz 11.25 kHz).

Die Signalanalyse mittels FFT-Transformation unter Verwendung einer Fensterfunktion ist in der Signalverarbeitung üblich. Gegenstand der Optimierung ist neben dem Frequenzband in erster Linie die Framebreite. Eine kleine Framebreite ist für eine gute Signalabbildung zwar wünschenswert, führt aber zu einer höheren Belastung der meist kritischen Prozessorkapazität. Für die weitere Verarbeitung wird das Pegelspektrum genutzt, das zum Ausgleich von Lautstärkeschwankungen in geeigneter Weise normiert wird. Durch eine Pegelbegrenzung nach unten lassen sich Störungen und Quantisierungsfehler ausblenden. Weiterhin erfolgte eine Begrenzung des Frequenzbereichs.

Die Reduktion des Spektrums auf 6 Koeffizienten pro Frame erfolgte über eine Hauptachsentransformation.

Die in die Datenbank aufgenommenen Mustersequenzen wurden als Indexsequenzen abgelegt, die mit Hilfe des zuvor bestimmten Codebuchs (2000..3000 Vektoren) ermittelt werden. Die zu verarbeitenden Input-Vektoren werden jedoch nicht quantisiert. Neben der Indexsequenz der Codebuch-Vektoren werden noch musterspezifische Schwellwerte (pro Frame) ermittelt. Der lokale Vergleich eines Input-Vektors mit einem Codebuchvektor beruht auf der Euklidischen Distanz.

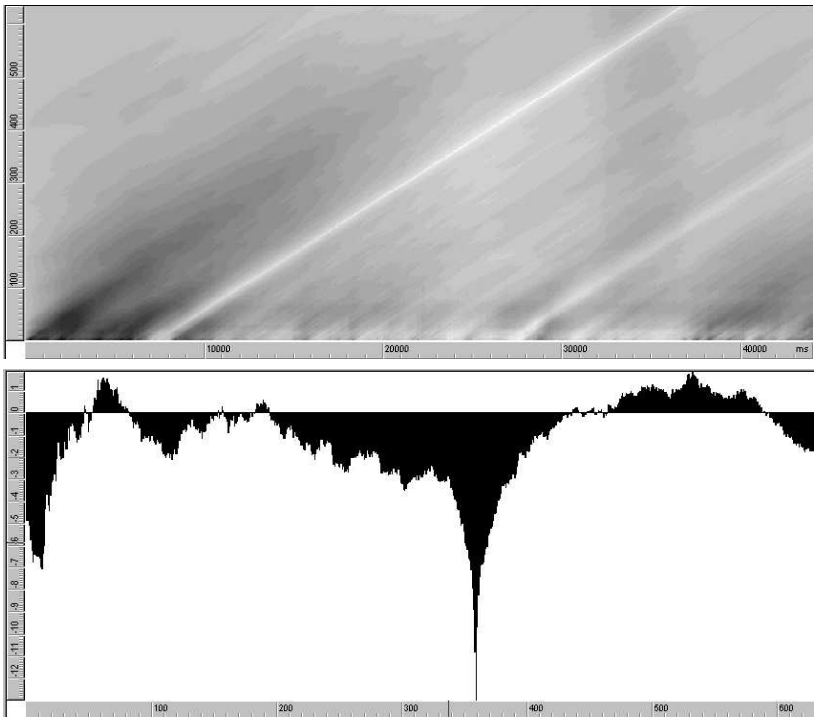

Abbildung 1 - oben: Verlauf der additiven normierten Bewertung einer Hypothese über der Zeit (x-Achse); unten: Querschnitt (Bewertungsprofil) für einen Zeitpunkt.

Der Mustervergleich von Sequenzen basiert in den meisten Fällen auf Verfahren der dynamischen Optimierung, die zum Ausgleich von Einfügungen, Auslassungen bzw. zeitlichen Verzerrungen dient. In den vorliegenden Fällen sind diese Verzerrungen allerdings von nur geringer Bedeutung und man kann demzufolge von einer elementeweisen Zuordnung

der Sequenzen ausgehen. Um die Auswirkung von Zeitverzerrungen weiter zu verringern, können die zu erkennenden Muster in kurze Abschnitte von 5 ... 10 s Länge zerlegt werden.

In Abbildung 1 ist dargestellt, wie sich ein Match zwischen einem Muster und einem Audiosignal darstellt. Die Pfadbewertung wird dabei auf die erwartete Distanz, die bei der Erstellung der Soundmuster mit bestimmt wird, normiert.

Wenn die additive normierte Bewertung nach ca. 80% der Gesamtlänge ein bestimmter Schwellwert erreicht, so wird eine entsprechende Hypothese generiert und in einem Hypothesenspeicher zwischengespeichert. Eine Kette von Hypothesen wird dann ausgegeben, wenn diese Kette einen gewissen Zeitraum überstreicht oder aber für eine gewisse Zeit keine Hypothese mehr detektiert wurde. Kriterien für die Zusammenfügung von Hypothesen zu einer Kette können sowohl zeitlicher (minimale Lücken/Überlappungen) als auch inhaltlicher Natur sein (z.B. durch Zerlegung eines Audiomusters entstanden).

4.2 Die Generierung der Datenbasis

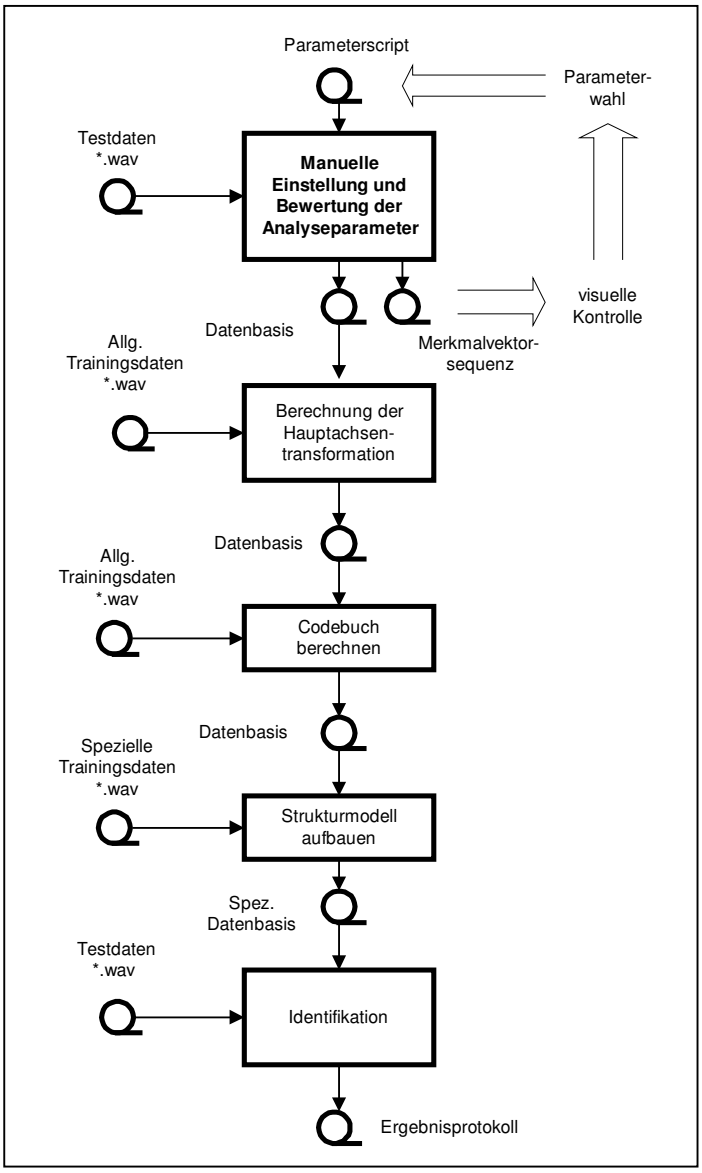

Abbildung 2 zeigt den Ablauf zur Erzeugung der Datenbasis. Zunächst werden die Signalanalyseparameter festgelegt. Da sich hier kaum objektive Optimierungsansätze bieten, erfolgte eine subjektive Kontrolle unter Verwendung verschiedener Audioproben. Gegenstand der Optimierung waren insbesondere die Zeitfensterbreite und die Fortsetzrate.

Nachdem die Analyseparameter festgelegt wurden, kann eine Hauptachsentransformation zur Merkmalsreduktion und anschließend ein Codebuch berechnet werden. Hierbei sollte möglichst viel unterschiedliches, aber typisches Datenmaterial verwendet werden.

Alle Parameter, die Matrix der Hauptachsentransformation und das Codebuch werden in einer Datenbank-Datei zusammengefasst, die dann die Grundlage der Generierung der spezifischen Datenbasis bildet bzw. in diese eingeht.

Abbildung 2 - Ablauf Generierung Datenbasis und Test

Die spezifische Datenbasis wird aus den Audiomustern gebildet, die später erkannt werden sollen. Diese Muster werden wie in der Erkennungsphase analysiert, transformiert und vektor-

quantisiert. Die entstehenden Indexsequenzen sowie die Distanzen gehen in Datenbasis ein. Sie bilden dann die Grundlage des oben beschriebenen Erkennungsvorgangs.

Die Datenbasis kann beliebig erweitert oder reduziert werden. An die Audiomuster werden bestimmte Anforderungen gestellt. Sie müssen eine Mindestlänge von 5 Sekunden aufweisen und es muss eine bestimmte Anzahl von Frames einen vorgebbaren Minimalpegel überschreiten. Bei sehr langen Mustern ist es empfehlenswert, diese in kleinere Teilabschnitte zu zerlegen, um eine merkliche Drift zwischen Original und Prototyp zu vermeiden.

5 Systemaufbau und Betriebserfahrungen

Abbildung 3 zeigt die Struktur eines Systems zur online-Überwachung des Programms von TV- oder Radiostationen. Pro Kanal benötigt man einen PC (Mainboard mit Audioeingang und Netzwerkanschluss) sowie eine Festplatte und eine TV-Karte. Der Audioausgang der TV-Karte wird mit dem Audioeingang des PC verbunden, das Videosignal wird nicht benötigt.

Ein weiterer Komplett-PC wird zur Administration des Systems benötigt. Auf diesem PC wird die Datenbasis aktualisiert und z.B. einmal täglich an die Kanal-PC übertragen. Weiterhin müssen die Ergebnislisten von den Kanal-PC abgeholt und kundenspezifisch sortiert werden. Schließlich sind die Verzeichnisse, in denen Signalmitschnitte zu Beweiszwecken erfolgen, zu sichern bzw. zu leeren, um einen Überlauf der Festplatte zu vermeiden.

Das implementierte Verfahren wurde im 24-Stunden-Betrieb Überwachung von ca. 20 Fernsehprogrammen eingesetzt, wobei es um die Erkennung von Werbespots ging. Das Verfahren arbeitete sehr stabil, es kam nur in komplizierteren Fällen zu Fehlerkennungen (bei Werbespots, die aus kurzen Audiobausteinen zusammengesetzt sind). Eine Neuberechnung des Codebuchs war nicht erforderlich.

Eine Hochrechnung hat ergeben, dass auf einem 2-GHz-PC mit dem Betriebssystem Windows-2000 eine Datenbasis, die ca. 25 h Audiomaterial enthält, noch in Echtzeit betreibbar ist. Je Sekunde Datenmaterial werden ca. 100 Byte benötigt. Dieser relativ hohe Speicherbedarf ist insbesondere auf die mitgeführten framespezifischen Schwellwerte zurückzuführen.

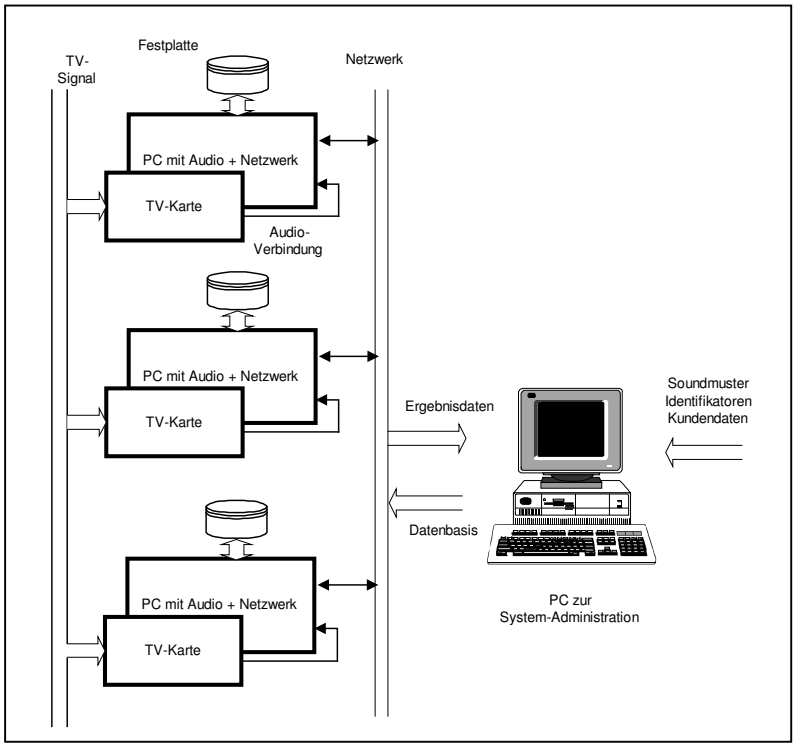

Abbildung 3 - Struktur eines Systems zur online-Überwachung von TV-Sendern

Literatur

[1] Der Sinn der Musik. Technology Review. Heise Verlag 2004. Im Internet http://www.heise.de/tr/artikel/44683

[2] Battle, E. et.al.: Automatic Song Identification in Noisy Broadcast Audio. In: Proc. of the SIP, Aug. 2002

[3] Battle, E. und P. Cano: Automatic Segmentation for Music Classification using Competive Hidden Markov Models. In: Proceedings Of International Symposium on Music Information Retrieval. 2000.

[4] Battle, E., J. Masip und E.Guaus: Amadeus: A Scalable HMM-based Audio Information retrieval System. Proc. of First International Symposium on Control, Communications and Signal Processing. Hammamet, Tunisia.

[5] Cano,P., O. M. Kaltenbrunner und E. Battle: Statistical Significance in Song-Spotting in Audio. Proceedings of International Symposium an Music Retrieval 2001. Bloomington, Indiana (USA).

[6] Cano,P., E. Batlle, H. Mayer,H. und H. Neuschmied: Robust Sound Modeling for Song Detection in Broadcast Audio. 112th Convention of the AES, 10.-13. Mai 2002, München

[7] Cano, P., E. Batlle, T. Kalker und J. Haitsma: A review of algorithms for audio fingerprinting. In International Workshop on Multimedia Signal Processing, US Virgin Islands, December 2002.

[8] Cano, P., E. Batlle, H. Mayer und H. Neuschmied: Robust sound modeling for song detection in broadcast audio. In Proc. AES 112th Int. Conv., Munich, Germany, May 2002.

[9] Gomes,L. et.al.: Audio Watermarking and Fingerprinting: For which Applications ? Journal of New Music Research. Vol. 32.1

[10] Haitsma, J. and Ton Kalker: A Highly Robust Audio Fingerprinting System. Proceedings of ISMIR 2002, Paris, France, October 2002.

[11] Kalker, T: Applications and Challenges for Audio Fingerprinting. Presentation at the 111th AES Convention, NY, in the "Watermarking versus Fingerprinting" workshop, December 3, 2001.