(19) österreichisches patentamt

(10) **AT 507 588 B1 2011-12-15**

(12) **Patentschrift**

(21) Anmeldenummer: A 9106/2008
(86) PCT-Anmeldenummer PCT/AT2008/000067
(22) Anmeldetag: 28.02.2008
(45) Veröffentlicht am: 15.12.2011

(51) Int. Cl. : ***G11B 20/00*** (2006.01)
G10L 21/04 (2006.01)

(30) Priorität:
08.03.2007 US 715766 beansprucht.

(56) Entgegenhaltungen:
US 2004068412A1
US 2006167688A1

(73) Patentinhaber:
UNIVERSITÄT FÜR MUSIK UND DARSTELLENDE KUNST
A-8010 GRAZ (AT)

(72) Erfinder:
HÖLDRICH ROBERT
GRAZ (AT)

(54) **VERFAHREN ZUM BEARBEITEN VON AUDIO-DATEN IN EINE VERDICHTETE VERSION**

(57) Aufgezeichnete Audiodaten (s1) werden komprimiert, um eine verdichtete Version (s2) zu erhalten, indem zuerst eine Anzahl von aufeinander folgenden, überlappungsfreien Segmenten (Block(k)) der Audiodaten ausgewählt wird, sodann jedes Segment durch zeitliche Kompression reduziert wird und die reduzierten Segmente in eine gekürzte Version kombiniert werden, die ausgegeben werden kann. Der Schritt der Audiodaten-Segmentierung besteht darin, ein Innovationssignal aus den Audiodaten abzuleiten, wobei das Innovationssignal eine Größe darstellt, die eine Änderungsrate des Inhalts in den Audiodaten angibt, Zeitpunkte von Maxima des Innovationssignals zu bestimmen, diese Zeitpunkte durch jeweilige Zeitversetzungen zu reduzieren und Segmentbeginnzeiten an den so reduzierten Zeitpunkten zu setzen.

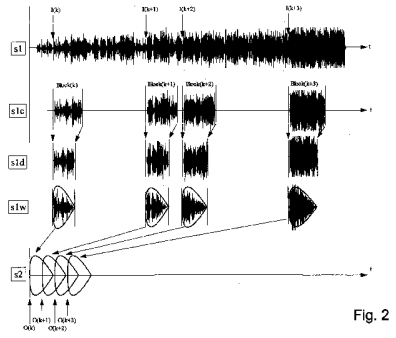

Fig. 2

AT 507 588 B1 2011-12-15

Beschreibung

VERFAHREN ZUM BEARBEITEN VON AUDIO-DATEN IN EINE VERDICHTETE VERSION
GEBIET DER ERFINDUNG UND BESCHREIBUNG DES STANDS DER TECHNIK

[0001] Die Erfindung betrifft ein verbessertes Verfahren zum Bearbeiten von in einer Aufnahme enthaltenen Audio-Daten, um eine gekürzte („verdichtete") Version zu erhalten, die zum Anhören (hörbar) wiedergegeben werden kann. Die Erfindung beinhaltet auch ein Verfahren zum Bearbeiten von Audio-Daten, um eine graphisch wiedergebbare Version zu erhalten.

[0002] Die Archive in Museen, Universitäten und anderen Institutionen führen ein kulturelles Vermächtnis von Millionen von Stunden von Audio-Video-Materialien (AVM), die auf Medien gespeichert sind. Große Teile dieser AVM sind nicht mit Annotationen versehen. Um ein systematisches Zugreifen und Erfassen dieser AVM zu gestatten, werden zeitsynchrone Metadaten hinzugefügt. Es ist schwierig und fehleranfällig, diesen Vorgang zu automatisieren, und Fehler müssen dann von Hand korrigiert werden. Zum Zwecke der Korrektur und Überprüfung muss der Benutzer schnell einen Überblick des vorliegenden AVM bekommen. Anders als bei Videomaterial, bei dem eine Übersicht durch Zusammenstellen einer Anzahl von Standbildern aus verschiedenen Epochen des Materials erstellt werden kann, ist es nicht sinnvoll oder überhaupt nicht möglich, eine bedeutungsvolle Kurzdarstellung des Audiomaterials in AVM zu erzeugen, die nicht eine gewisse Bearbeitung in ablaufender Zeit vorgesehen ist.

[0003] Untersuchungen von AVM, wie z.B. Studien über die Verwendbarkeit von Bildschirmlesegeräten bei sehbehinderten Personen, zeigten dass die beschleunigte Wiedergabe von Sprache die Verständlichkeit bereits bei einem Beschleunigungsfaktor von 2-3 bedeutend verringert, sogar für trainierte Benutzer. Mit Beschleunigungsfaktoren, die geringfügig höher sind (max. 4-6), ist es möglich, ein Musikstück zu erkennen, wenn es sich um bestimmte Arten von Liedern handelt. In diesen beiden Beispielen wurde reine Zeitkompression ohne Tonhöhenverschiebung verwendet.

[0004] Bekannte Verfahren zur beschleunigten Wiedergabe von Audiomaterialien zielen hauptsächlich auf Sprache (gesprochene Worte) ab, wobei die völlige Verständlichkeit des Textes im Vordergrund steht. Das System „Speechskimmer" wird von B. Arons in 'SpeechSkimmer: A System for Interactively Skimming Recorded Speech' („Speechskimmer: ein System zum interaktiven Skimmen von Sprachaufnahmen") - ACM Transactions on Computer-Human Interaction, Vol. 4, Nr. 1, S. 3-38, 1997, beschrieben. Es verwendet Zeitkompressionsverfahren, wie z.B. das SOLA-Verfahren ('Synchronized OverLap Add', etwa: synchronisiertes Überlappen und Zusammensetzen), dichotisches Sampling (was eine binaurale Wiedergebe erfordert) oder Extraktion von Pausen und Skimming-Techniken, die Teile des Sprachsignals auslassen. Isochrone Verfahren geben feste Zeitsegmente wieder, die aus dem gesamten Signal ausgeschnitten worden sind (z.B. die ersten fünf Sekunden jeder einminütigen Zeitdauer); sprachsynchrone Verfahren wählen wiederzugebende Segmente durch Aufteilen des Sprachsignals in wichtige und weniger wichtige Teile aus, auf Grundlage von Charakteristika wie z.B. Pausendetektion, Leistungs- und Tonhöhenverlauf, eine Sprechererkennung und Kombinationen von diesen. Eine anderes Verfahren zum Segmentieren, das von D. Kimber und L. Wilcox in 'Acoustic segmentation for audio browsers' („Akustische Segmentierung für Audio-Browser") - Proc. Interface Conference, Sydney, Australia, 1996, verwendet Hidden-Markov-Modelle. Das von S. Lee und H. Kim in 'Variable Time-Scale Modification of Speech Using Transient Information' („Variable Sprachmodifizierung in der Zeitskala mittels transienter Information") - 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'97), Vol. 2, S. 1319-1322, 1997, beschriebene Verfahren belässt die Sprachübergänge ungeändert und komprimiert nur die stationären Komponenten wie Vokale, wodurch eine bessere Verständlichkeit der Sprache erreicht wird. Alle diese Verfahren sind auf Sprachinhalte eingeschränkt und erzeugen keine guten Ergebnisse für Audiomaterialien, die andere Inhalte wie z.B. Musik oder Hintergrundgeräusche enthalten.

[0005] Ein Verfahren zum Skimmen digitaler Audio- und Videodaten ist in der WO 96/12240

beschrieben.

[0006] Gupta, in US 7,076,535, und N. Omoigui et al. in 'Time-Compression: System Concerns, Usage, and benefits' („Zeitkompression; Systemforderungen, Anwendung und Nutzen") - Proceedings der SIGCHI Conference on Human Factors in Computing Systems, S. 136-143, ACM Press, 1999, beschreiben eine Client-Server-Architektur zum Skimmen von Multimedia-Daten, gehen jedoch nicht auf die tatsächlich verwendeten Verfahren außer dem bereits erwähnten SOLA-Verfahren ein.

[0007] In US 2004/0068412 A1 ist ein Verfahren einer Energie-basierten, nicht-uniformen zeitlichen Kompression von Audiosignalen beschrieben, bei dem die erhaltenen Daten in Segmente aufgegliedert werden und für jedes Segment der Energiegehalt bestimmt wird; aufgrund des Energiegehalts wird die zeitliche Kompression der Audidaten gesteuert.

[0008] Die US 2006/0167688 A1 offenbart die Kompression von Audiosignalen und Multimedia-Daten mittels eines Lempel-Ziv-Algorithmus.

[0009] Weitere Verfahren zur Bearbeitung von Audioinformation sind in den beiden Artikeln von G. Tzanetakis und P. Cook, "Multifeature audio segmentation for browsing and annotation", IEEE Workshop on Applications of Signal Processing to Audio and Acoustices, NY, USA, 1999, Seiten 103-106 und "Audio Information Retrieval (AIR) Tools", Proc. Internationale Symposium on Music Information Retrieval (ISMIR), Plymouth, MA USA, 2000, beschrieben.

KURZFASSUNG DER ERFINDUNG

[0010] Die Erfindung sieht Umsetzungen einer Verdichtung von Audio-Daten in einer Weise vor, die keine vollständige Verständlichkeit der Sprache oder Erkennbarkeit einer musikalischen Komposition verlangt. Vielmehr soll es ausreichen, einen groben aber repräsentativen Überblick des vorliegenden Materials zu liefern. Die AVM-Arten sind nicht auf lediglich Sprache oder Musik beschränkt. Zudem sind Kompressionsfaktoren von bis zu 30 oder sogar mehr gewünscht.

[0011] Dieses Ziel wird von einem Verfahren zum Bearbeiten von in einer AVM-Aufnahme enthaltenen Audio-Daten zum Gewinnen einer zum Anhören wiedergebbaren gekürzten Version, mit den Schritten

[0012] - Auswahl einer Anzahl von aufeinander folgenden, nicht-überlappenden Segmenten der Audiodaten,

[0013] - Reduktion jedes Segments durch zeitliche Kompression, und

[0014] - Kombinieren der so reduzierten Segmente,

[0015] gelöst, wobei der Schritt der Audiodaten-Segmentierung die Teilschritte aufweist, ein Innovationssignal aus den Audiodaten abzuleiten, wobei das Innovationssignal eine Größe darstellt, die eine Änderungsrate des Inhalts in den Audiodaten angibt, Zeitpunkte von Maxima des Innovationssignals zu bestimmen, diese Zeitpunkte durch jeweilige Zeitversetzungen zu reduzieren und Segmentbeginnzeiten an den so reduzierten Zeitpunkten zu setzen.

[0016] Die Erfindung stellt ein Verfahren zur Verfügung, welches das Erstellen einer - je nach Wunsch - zum Anhören und/oder Ansehen abspielbaren verdichteten Darstellung großer Audio- und AVM-Dateien (nämlich mit einer Dauer von mehreren Minuten bis zu einigen Stunden) mit einem großen Gesamtkompaktierungsfaktor ermöglicht.

[0017] Das erfindungsgemäße Verfahren ist nicht auf Sprachinhalte beschränkt. Obwohl die Zeitkompressionsalgorithmen des SpeechSkimmer ähnlich sein mögen, sind die zur Auswahl der Segmente verwendeten Skimming-Verfahren allgemeiner und beruhen auf dem Leistungsverlauf des Signals, welches auf verschiedene Arten spektral gewichtet wird, um signifikante Änderungen der Signalcharakteristik zu detektieren. Außerdem werden die Segmente überlappt, um mehrere Segmente zur gleichen Zeit hörbar zu machen. Das ist in markantem Gegensatz zum SOLA-Verfahren, das Segmentlängen und Überlappungen im Bereich von einigen

wenigen 10 ms verwendet.

[0018] In einer Weiterbildung der Erfindung wird die zeitliche Kompression mit einem lokalen Kompressionsfaktor ausgeführt, der zwischen den Segmenten variiert. In einem Spezialfall, der zum Herausheben eines zentralen Fokus des Audiomaterials dient, kann der lokale Kompressionsfaktor einen Minimalwert (der lediglich 1 betragen kann, d.h. keine wirkliche Kompression) für ein Mittelsegment annehmen. Außerdem kann der lokale Kompressionsfaktor über die Segmente vor diesem Mittelsegment insgesamt abnehmen und über die Segmente nach diesem Mittelsegment insgesamt zunehmen.

[0019] Verschiedene bevorzugte Verfahren zum Ableiten eines solchen Analysesignals, auch als Innovationssignal bezeichnet, werden in der Beschreibung weiter unten diskutiert. Beispielsweise kann es zweckmäßig sein, eine Aufteilung eines Audiodaten-Signals in eine Anzahl von Frequenzband-Signalen, eine Berechnung einer entsprechenden Zahl sekundärer Signale aus den Frequenzband-Signalen mithilfe zumindest eines der folgenden Verfahren: Filtern des Signals, Glätten des Signals und Berechnen eines lokalen Polynoms aus dem Signal; dann Kombinieren der sekundären Signale in einen mehrdimensionalen Leistungsvektor P(n) und eine Berechnung einer Distanzfunktion zwischen dem aktuellen und einem vorangehenden Wert des Leistungsvektors zum Bilden des Innovationssignals Inno(n) = dist[P(n) - P(n-m)] durchzuführen.

[0020] Ein anderes geeignetes Verfahren zur Berechnung des Innovationssignals verwendet Meta-Merkmal-Vektoren. Ein zweckmäßiger Weg zum Berechnen der Meta-Merkmal-Vektoren besteht darin, die Segmente der Audiodaten in Untersegmente aufzuteilen, Merkmalsvektoren für diese Untersegmente zu berechnen, Verteilungsparameter dieser Merkmalsvektoren zu berechnen, und diese Verteilungsparameter in einen Meta-Merkmal-Vektor zu kombinieren. Das Innovationssignal wird berechnet durch Segmentieren der Audiodaten in überlappungsfreie Segmente, Berechnen eines Meta-Merkmal-Vektors F(l) aus jedem dieser Segmente, Durchführen einer k-Means-Clusteranalyse für die so erhaltenen Meta-Merkmal-Vektoren und, um das Innovationssignal zu erhalten, Berechnen eines Markersignals für jedes Segment durch Zuweisen eines positiven Werts dann, wenn der Meta-Merkmal-Vektor in einem von dem Cluster des vorangehenden Segments verschiedenen Cluster liegt, bzw. eines Wertes Null ansonsten. Die k-Means-Clusteranalyse kann mehrfach vorgenommen werden, nämlich für G verschiedene Werte der Zahl k_g der Cluster mit g=1,...,G, wobei G Markersignale für jedes Segment erhalten werden; das Innovationssignal kann dann durch Mittelung einer Überlagerung dieser Markersignale $Mark_g$ unter Verwendung einer Glättungsfunktion $A\mathcal{V}$ berechnet werden, um das Innovationssignal $Inno(l) = A\mathcal{V}(\sum_g Mark_g(l))$ zu erhalten. Nähere Einzelheiten dieses Berechnungsverfahrens sind in der Beschreibung detailliert beschrieben.

[0021] Die Segmentierung der Audiodaten kann beruhend auf Nicht-Audio-Daten ausgeführt werden, die in der Aufnahme enthalten und zudem synchron mit den Audiodaten sind. In diesem Fall können die Segmentbeginnzeiten bei in den Nicht-Audio-Daten vorhandenen Zeitmarkierungen platziert werden.

[0022] Ein einfaches Vorgehen zum Kombinieren der reduzierten Segmente ist, sie in chronologischer Reihenfolge in Bezug auf ihre ursprüngliche Position in den Audiodaten zusammenzufügen, unter Auswahl entweder der voran- oder der rückwärtslaufenden Ordnung.

[0023] Eine zusätzlich Kompaktierung der Audiodaten kann erreicht werden, wenn der Schritt des Kombinierens der reduzierten Segmente eine Überlagerung der Segmente enthält. Dies kann eine gestaffelte Überlagerung sein, wobei die Segmente zu aufeinander folgenden Beginnzeiten anfangen und jedes nicht-erste Segment eine Beginnzeit innerhalb der Dauer des jeweils vorangehenden Segments hat.

[0024] Beruhend auf den vorangehend beschriebenen Verfahren stellt die Erfindung auch ein Verfahren zum Bearbeiten von Audio-Daten zum Gewinnen einer graphisch darstellbaren Version zur Verfügung, enthaltend die Schritte:

[0025] Ableiten eines Analysesignals aus den Audiodaten, wobei das Analysesignal eine Größe

darstellt, die eine Änderungsrate des Inhalts in den Audiodaten angibt (das Analysesignal kann durch eines der hier beschriebenen Innovationssignalverfahren abgeleitet werden), Bestimmen von Zeitpunkten von Maxima des Innovationssignals, Setzen von Segmentgrenzen an so reduzierten Zeitpunkten und Anzeigen der so definierten Segmente in einer linearen Abfolge von Flächen mit variierender graphischer Wiedergabe.

[0026] Es ist einzusehen, dass die oben erwähnten und in den abhängigen Ansprüchen beschriebenen Weiterbildungen der Erfindung nicht voneinander getrennt zu sehen sind, sondern miteinander kombinierbar sind.

KURZBESCHREIBUNG DER ZEICHNUNGEN

[0027] Im Folgenden wird die Erfindung in näheren Einzelheiten unter Bezugnahme auf die Zeichnungen beschrieben, welche zeigen:

[0028] Fig. 1 ein Blockdiagramm-Schema einer Implementation der Erfindung, welches ein Kompressionmodul enthält;

[0029] Fig. 2 das Funktionsprinzip des Kompressionsmoduls;

[0030] Fig. 3 illustriert die Verwendung eines Innovationssignals zum Festlegen einer Segmentgrenze; und

[0031] Fig. 4 ein Beispiel einer graphischen Darstellung von Audiodaten.

AUSFÜHRLICHE BESCHREIBUNG DER ERFINDUNG

KOMPRESSIONSENGINE

[0032] Fig. 1 zeigt ein schematisches Blockdiagramm einer Umsetzung des Verfahrens gemäß einer beispielhaften Ausführungsform der Erfindung. Die auch als AudioShrink bezeichnete Umsetzung kann als eine Einrichtung 100, z.B. ein Computersystem, ausgebildet sein. Es weist eine Anzahl von Funktionsblöcken wie folgt auf. Ein erster Funktionsblock FB1 liest Audiodateien als Audio-Eingangssignal 1 ein. In der gezeigten Ausführungsform ist er mithilfe einer Festplatte oder einem anderen Permanentspeicher realisiert, auf der/dem Audiodateien gespeichert werden. Eine andere mögliche Ausbildung des Blocks FB1 ist eine Schnittstelle für den Zugriff auf und Abruf von Audiodaten, beispielsweise über das Internet. Der Block FB1 kann fehlen, wenn der Einrichtung die Audioeingabe direkt in der passenden elektrischen Signalform geliefert wird. Ein zweiter Funktionsblock FB2 ist ein Kompressionsmodul, das das Audiomaterial 1 von dem Block FB1 empfängt und eine Zeitkompression durchführt, um eine komprimierte Audioausgabe 2 zu erzeugen. Das Kompressionsmodul FB2 kann mehrstufig sein; es ist weiter unten ausführlicher beschrieben. Ein dritter Funktionsblock FB3 spielt die Audioausgabe 2 durch Erzeugen eines hörbaren (oder auf andere Art wahrnehmbaren) Signals 3 ab. Der Block FB3 ist beispielsweise mitthilfe einer Computer-Soundkarte mit einem Digital-Analog-Konverter realisiert, der mit geeigneten Schallerzeugergeräten wie Lautsprechern oder einem Kopfhörergerät verbunden ist. Ein vierter Funktionsblock FB4 dient als Steuermodul, das die mehrstufige Kompression im Block FB2 durch Steuerparameter 4 wie weiter unten beschrieben steuert.

[0033] Außerdem kann wahlweise ein fünfter Block FB5 vorgesehen sein, der das von Block FB1 gelieferte Audiomaterial analysiert und Analyseresultate erzeugt, in Form eines Analysesignals 5, als Eingabe für den Steuerblock FB4 zusätzlich zu externen Eingaben, die von dem Benutzer eingegeben werden, wie z.B. einem gewünschten Kompressionsfaktor 5b oder Befehle 5c, nach vorne oder zurück zu springen. Zudem kann das Analysesignal 5 für eine graphische Darstellung der Struktur des Audiosignals 1 verwendet werden.

[0034] Es ist anzumerken, dass im Rahmen dieser Offenbarung der Begriff Kompression sich auf eine zeitliche Kompression (also mit einer kürzeren Zeitdauer) bezieht. Dies ist nicht mit einer dynamischen Kompression des Audiomaterials zu verwechseln.

BEI DER KOMPRESSION EINGESETZTE VERFAHREN

[0035] Die zeitliche Kompression wird an der gesamten Audiodatei, die dem Kompressionsmodul (Funktionsblock FB2) übergeben wird, durchgeführt. Drei miteinander kombinierbare Stufen sind implementiert: (1) reine zeitliche Verkürzung, (2) Überlagerung (Superposition) und (3) Auswahl.

[0036] 1) Reine zeitliche Verkürzung: Der Begriff reine zeitliche Verkürzung soll hier ein zeitliches Stauchen ('Squeeze', beschleunigte Wiedergabe) bezeichnen, das von einer Verschiebung der Tonhöhe begleitet sein oder ohne diese erfolgen kann. Dies kann mit bekannten Verfahren wie Variable-Speed-Replay (Abspielen mit variabler Geschwindigkeit) oder Granularsynthese erfolgen. Korrelationsbasierte Verfahren können auch verwendet werden, wie z.B. synchrones Overlap-and-Add (Überlappen und Zusammenfügen) oder - besonders für Sprache - Tonhöhen-synchrones Overlap-and-Add. Außerdem können den Frequenzbereich erhaltende Techniken, wie z.B. Sprach-Vocoder, geeignet sein. Zusätzlich zur eigentlichen Zeitkompression kann eine Tonhöhentransposition eingerichtet sein. Eine reine zeitliche Verkürzung erbringt typischer Weise Kompressionsfaktoren von 2 bis 4.

[0037] 2) Überlagerung: Dies ist das gleichzeitige Ablaufen mehrerer Segmente mit oder ohne wechselnden räumlichen Bedingungen (im Falle stereophonischer oder anderer räumlicher Darbietung). Dieser Aspekt nutzt die Fähigkeit des menschlichen Ohrs aus, Information aus akustischer Information zu extrahieren, die in denselben oder überlappenden Intervallen gespielt wird. Das Audiosignal wird in eine Anzahl angrenzender Segmente aufgeteilt, die superponiert (überlagert) werden, sodass sie zur selben Zeit gespielt werden. Beispielsweise kann ein Audiomaterial von 60 s durch eine 4fache Überlagerung in 15 s umgewandelt werden. Um ein Trennen der überlagerten Ebenen zu unterstützen, kann ein räumlicher Ablauf hinzugefügt werden, wie z.B. Ausgabe des Beginns des Segments über den linken Kanal und kontinuierliches Schwenken zum rechten Kanal bei Segmentende („vorbei fahrendes Fahrzeug").

[0038] 3) Auswahl (Fortlassung): Nur ausgewählte Segmente des Materials werden verarbeitet; die übrigen Teile werden übersprungen. Die Länge der beibehaltenen Segmente wird in geeigneter Weise gewählt, so dass ein Erkennen des Inhalts des einzelnen Segments möglich bleibt, während eine ausreichende Homogenität zwischen benachbarten zu spielenden Segmenten gesichert ist, um eine kategoriale Änderung in den Audiosegmenten transparent zu machen. Die Auswahl von zu behaltenden Audiosegmenten (im Gegensatz zu auszulassenden Segmenten) kann aufgrund einer vom Benutzer gelieferten Parameter-Auswahl (feste Parameter) und/oder aufgrund von Analyseparametern (dynamische Auswahl) stattfinden, die den Analyseergebnissen 5 des Analysemoduls FB5 entnommen wurden, oder - im Falle audiovisueller oder anderer kombinierter Daten - Information, die von dem Video bzw. anderen nichtakustischen Daten abgeleitet wurde. Es wird erwartet, dass die auswählende Darstellung eine Kompression von zwischen 3 und 6 bei festen Parametern ergibt, während Faktoren von ca. 20 oder mehr mit dynamischer Auswahl erzielbar sind.

[0039] Die obigen Kompressionsverfahren können kombiniert werden. Beispielsweise kann eine Kombination von reiner zeitlicher Verkürzung und Überlagerung verschiedener Audiosegmente gemacht werden. In diesem Fall kann eine zeitlich variierende Tonhöhenverschiebung jedes Segmentes die Erkennbarkeit der Segmentinhalte verbessern. Die Tonhöhenverschiebung kann z.B. von einer Tonerhöhung am Segmentbeginn zu einer Tonerniedrigung am Segmentende übergehen.

STEUERUNG DER KOMPRESSION

[0040] Der Funktionsblock FB4 ist das Steuermodul zum Steuern der mehrstufigen zeitlichen Kompression. Eine Kombination der oben diskutierten Kompressionsstufen gestattet die Kompaktierung von Audiomaterial um einen Faktor von bis zu 50 oder sogar mehr. Das bedeutet, dass z.B. eine 5-Minuten-Sequenz in 6 s dargebracht werden kann, oder ein Schnelldurchlauf durch ein einstündiges Audiomaterial nur 1 bis 2 Minuten braucht. Das Steuermodul setzt den Gesamtkompressionsfaktor und die Wiedergaberichtung (vorwärts oder rückwärts) gemäß den

Benutzereingaben. Außerdem setzt es eine Kombination der Kompressionsstufen (1) bis (3) mit einzelnen Kompressionsfaktoren, um den Gesamtkompressionsfaktor zu erhalten. Das Steuermodul interagiert auch mit dem Benutzer und erhält und interpretiert gegebenenfalls das Analysesignal 5 von dem Analysemodul FB5.

[0041] Das Analysemodul FB5 liefert Information zum Auswählen relevanter Teile des Audiomaterials, durch Ausgabe dieser Information in Form eines Analysesignals 5. Das Hauptpotenzial der zeitlichen Kompression liegt in der selektiven Darstellung von Audiomaterial, d.h. Fortlassung von Teilen. Neben einer festen Aufteilung in darzustellende und wegzulassende Segmente - beispielsweise eine Segmentierung in 2,5 s-Teile, zwischen denen 5 s weggelassen werden, was einen Kompressionsfaktor 3 ergibt - sind zweckmäßige Verfahren solche, die „relevante" Audioinformation finden, während weniger wichtige oder redundante Teile unterdrückt werden. Die folgenden Fälle sind beachtenswert:

[0042] a) Verfahren, die auf Audiomaterial-Analyse beruhen

[0043] Die Audioinformation kann in ein „Innovationssignal" umgearbeitet werden, das die Audioinformation charakterisiert - in dem Sinne, dass eine (ausreichend erhebliche) Änderung des Innovationssignals den Anfang eines Abschnitts mit neuen Inhalten oder neuen Kennzeichen anzeigt -, und dieses Innovationssignal kann als Analysesignal 5 zusammen mit einer passenden Heuristik des Steuermoduls FB4 verwendet werden. Das Innovationssignal kann mithilfe bekannter Signalverarbeitungsverfahren aus den Gebieten des Audio-Information Retrieval („Audioinformationsabfrage"), Signalklassifizierung, Ansatz- oder Rhythmus-Detektion, Voic-Activity Detection („Stimmenaktivitätsdetektion") oder anderen, sowie geeignete Kombinationen von diesen, bestimmt werden. Das Ergebnis einer derartigen Analyse kann eine Menge von Markerpunkten beinhalten, die den Beginn verschiedener Abschnitte und wiederum Relevanzinformation für die Charakterisierung anzeigen.

[0044] Ein im AudioShrink verwendeter Algorithmus von besonderem Interesse ist ein Verfahren, das auf einem fortschreitenden Multilevel- („Mehrfachrunden"-) k-Means-Clustering von Merkmalsvektoren, wie z.B. mel-Frequenz-Cepstrumkoeffizienten, beruht. Um die Dimension der eingesetzten Merkmalsvektoren zu verringern, kann eine Hauptkomponentenanalyse verwendet werden. Die Ergebnisse dieses Verfahrens eignen sich auch für eine graphische Darstellung von Audiomaterial (siehe unten). Das im AudiShrink verwendete Verfahren ist eine Erweiterung des Verfahrens, das von G. Tzanetakis und P. Cook in '3d Graphics Tools for Sound Collections', Proc. Conference on Digital Audio Effects, Verona, Italien 2000, zur Erzeugung von „Timbregrammen" präsentiert wurde. Im Gegensatz zu Tzanetakis funktioniert Clustering im Rahmen des AudioShrink mit einem fortschreitenden k-Means-Algorithmus (anstatt einem k-Nächste-Nachbarn-Algorithmus) und wird in mehreren Levern („Runden") ausgeführt. Somit wird in Abhängigkeit von dem Kompressionsfaktor der akustischen/graphischen Darstellung eine wechselnde Zahl von Klassen verwendet, und folglich von zu einer Klasse gehörenden Segmenten wechselnder Länge. Selbstverständlich können ebenfalls andere Algorithmen zum Ableiten eines Innovationssignals geeignet sein.

[0045] b) Verfahren, die Information aus Video- oder Meta-Daten nutzen

[0046] Falls das vorliegende Material auch synchrone Multimedia-Information umfasst, wie z.B. synchrone Mediadaten von Videomarkern, können diese Daten als Indikatoren für den Beginn einer Szene genutzt werden. Das Material, dass einem solchen Punkt unmittelbar zeitlich folgt, wird dann als relevant betrachtet und deshalb wird seine Wiedergabe bevorzugt.

KOMPRESSIONSMODUL - MEHRSTUFIGE VARIABLE KOMPRESSION

[0047] Fig. 2 stellt ein Beispiel dafür dar, wie eine Anzahl aufeinanderfolgender Signalverarbeitungsstufen zu einer mehrstufigen Kompression im Kompressionsmodul (Funktionsblock FB2) kombiniert sind. Die Wiedergaberichtung ist in dem gezeigten Beispiel „vorwärts". In Fig. 2 sind Audiosignale in Abhängigkeit von der Zeit t (horizontale Achse) in verschiedenen Schritten des mehrstufigen Vorgangs gezeigt; das oberste Signal gibt dabei das ursprüngliche Audiosignal s1 wieder. Das Signal s1 kann ein über die Zeit kontinuierliches Signal s1(t) sein, oder ein diskre-

tes Signal s1(n) zu diskreten Zeitpunkten, insbesondere bei einem digitalen Signal, wobei die Zeitspanne zwischen aufeinanderfolgenden Zeitpunkten n ausreichend klein ist, dass der Zuhörer das Signal s1 insgesamt als Kontinuum wahrnimmt.

[0048] Das Signal s1 füllt die in Fig. 2 gezeigte Zeitspanne weitgehend aus. Das Steuermodul FB4 bestimmt eine Anzahl von Auswahlpunkten I(k), k = 1,...,K. Jeder Auswahlpunkt I(k) stellt einen Zeitpunkt dar und gibt die Beginnzeit eines „relevanten" Signalblocks an. Da die Wiedergabe vorwärts ist, gilt I(k) > I(k-1) für alle Auswahlpunkte, (bei einer Rückwärtswiedergabe I(k) < I(k-1).) Die Gesamtzahl K der Blöcke hängt von dem Audiomaterial ab; im gezeigten Beispiel ist K = 4.

[0049] Die Blöcke Block(k) werden ausgehend von entsprechenden Auswahlpunkten I(k) mit einer gemeinsamen Länge N ausgewählt, wodurch sich ein zerteiltes Signal s1c ergibt. Die Blocklänge N wird ebenfalls von dem Steuermodul FB4 geliefert. Im allgemeinen wird die Länge N so gewählt, dass

$$N \leq N_{CF} + |I(k) - I(k-1)|,$$

[0050] wobei N_{CF} die Überblendelänge ist, d.h. die Dauer der für ein Überblenden benötigten Mindestüberlappung.

[0051] Dann wird jeder Block um einen Stauchungsfaktor C (rein zeitliche Verkürzung) komprimiert, unter Verwendung geeigneter Verfahren wie teilweise oder vollständige Reduktion von Pausen innerhalb eines Blockes, SOLA, Granularsynthese (asynchrones Overlap-and-Add), Phasenvocoder oder Resampling (einschließlich Tonhöhenverschiebung). Das so erhaltene Signal ist in Fig. 2 als s1d bezeichnet. Dann wird jeder Block gemäß einer Fensterlänge N_w und einer Fensterform, die von den Steuermodul FB4 bestimmt wurde, gefenstert. Die Fensterfunktion ist in Fig. 2 bei dem Signal s1w als eine jeden gefensterten Block umgebende Kontur dargestellt.

[0052] Schließlich werden die Blöcke Block(k) zu dem endgültigen AudioShrink-Signal s2 zusammengefügt (superponiert). Jeder Block wird zu einer Zeit bewegt, die durch vom Steuermodul ebenfalls gelieferte Beginnzeiten O(k) definiert sind.

[0053] Der Gesamtkompressionsfaktor C_{tot} entspricht dem Verhältnis zwischen dem mittleren Zeitabstand ΔI zwischen benachbarten Auswahlpunkten im ursprünglichen Signal und dem mittleren Zeitabstand ΔO zwischen benachbarten Blockanfängen im AudioShrink-Signal:

$$C_{tot} = \Delta I / \Delta O; \qquad \Delta I = (1/K) \Sigma_k (I(k) - I(k-1));$$
$$\Delta O = (1/K) \Sigma_k (O(k) - O(k-1));$$

[0054] Der mittlere Überlappfaktor Ovp im AudioShrink-Signal kann über $Ovp = N_W/\Delta O$ berechnet werden.

STEUERMODUL - BERECHNEN MEHRSTUFIGER KOMPRESSIONSPARAMETER

[0055] Die Steuerparameter der oben beschriebenen Kompression werden vom Funktionsblock FB4 geliefert, beruhend auf dem Gesamtkompressionsfaktor C_{tot}, der üblicherweise vom Benutzer vorgegeben wird. Üblicherweise ist C_{tot} eine Konstante, aber optional kann es ein zeitabhängiger Wert $C_{tot}(t)$ sein. Die Parameter sind: N - Länge der ausgewählten Blöcke; N_{CF} - Mindestüberlappung bei Überblenden; I(k) - Auswahlpunkte mit k=1...K; O(k) - Beginnzeiten mit k=1...K; C - Kompressionsfaktor; N_W - Fensterlänge; und die Fensterform, die z.B. über eine Funktion w(t) oder durch Angabe eines Typ-Index aus einem vorgegebenen Satz von Fensterform-Typen definiert werden kann. Im Allgemeinen kann die Beziehung zwischen den Steuerparametern und dem Gesamtkompressionsfaktor über eine Polynomfunktion oder mittels Nachschlagetabellen angegeben werden. Typische Werte der Parameter sind in Tabelle 1 wiedergegeben.

N_W = 3 bis 6 s;
N_{CF} = 30 bis 100 ms;
Fensterform =
Hanning, Dreieck, Tukey, oder Rechteck mit linearer Ein- und Ausblendung;
C = 1 bei C_{tot} = 1, linearer Anstieg bis
= 2 bei $C_{tot} \geq 20$;
N = $N_W C + N_{CF}$;
O(k) = $O(k-1) + N_W/C^2$;
I(k) = $I(k-1) + C_{tot}(O(k) - O(k-1)) = I(k-1) + N_W \cdot C_{tot}/C^2$;
k_1 = 2 bis 5.

Tabelle 1: Typische Werte von Kompressionparametern

[0056] Wenn ein Analysemodul FB5 zur Auswahl relevanter Audioinformation verwendet wird, ergibt die Signalanalyse Information für die Auswahl von Blöcken, die die isochrone Blockauswahl, d.h. die Wahl der Parameter I(k) und O(k), in Tabelle 1 ersetzt. Das Analysemodul FB5 erzeugt ein Innovationssignal Inno(t), das eine kontinuierliche oder diskrete Sequenz ist, die den Neuheitsgrad des ursprünglichen Audiosignals s1(t) angibt. Wenn ein Bereich im Signal einen hohen Innovationsgrad hat, besteht eine höhere Wahrscheinlichkeit, dass dieser Bereich ausgewählt und dann ein Auswahlpunkt I(k) entsprechend gesetzt wird. Das ergibt eine Integration der herausstechenden Klangsequenzen, d.h. sich von dem vorangehenden Material deutlich sich unterscheidenden Sequenzen, in das AudioShrink-Signal s2(t). Deshalb sind die Zeitabstände I(k) - I(k-1) zwischen zwei benachbarten Auswahlpunkten im Allgemeinen nicht für alle Werte von k gleich. Um den vorgeschriebenen Gesamtkompressionsfaktor C_{tot} einzuhalten, ist es wichtig, dass das Verhältnis zwischen dem mittleren Zeitabstand ΔI zwischen benachbarten Auswahlpunkten im ursprünglichen Signal und dem mittleren Zeitabstand ΔO zwischen benachbarten Blockanfängen eingestellt wird. Hierfür hat sich das folgende Vorgehen als zweckmäßig herausgestellt:

[0057] Wenn ein Auswahlpunkt I(k) ausgewählt wird, wird zuerst ein vorläufiger Wert $I_{target}(k)$ gemäß

$$I_{target}(k) = C_{tot} \cdot O(k);$$

berechnet. Im Falle einer zeitabhängigen Definition von $C_{tot}(t)$ wird der vorläufige Wert $I_{target}(k)$ über

$$I_{target}(k) = C_{tot} \cdot O(k) \text{ für } k \leq k_1;$$
$$I_{target}(k) = C_{tot}(t) \cdot [O(k)-O(k-k_1)]+I(k-k_1)$$

berechnet, wobei k_1 eine kleine ganze Zahl ist (typische Werte für k_1 sind in Tabelle 1 angegeben). Dieser vorläufige Wert ist die Zeit, die das gewünschte C_{tot} zusammen mit den anderen Parametern ergeben würde. Fig. 3 illustriert das Bestimmen des Auswahlpunktes I(k), ausgehend von einem vorläufigen Wert $I_{target}(k)$ für ein Signal s1(t) und einem daraus abgeleiteten Innovationssignal Inno(t). Das Innovationssignal wird mit einer bei $t_0=I_{target}(k)$ zentrierten Fensterfunktion $f(t-t_0)$ multipliziert. Die Fensterfunktion dient zum Herausprojizieren eines Abschnitts des Innovationssignals innerhalb einer endlichen Fensterdauer $2t_W$. In dem in Fig. 3 gezeigten Beispiel ist die Fensterfunktion eine Dreiecksfunktion, die mit unterbrochenen Linien dargestellt ist. Im Allgemeinen wird eine Fensterfunktion so gewählt, dass sie im Zentrum des Fensters den Wert 1 annimmt (d.h. $f(t-t_0=0) = 1$), für die Zeiten außerhalb des Zeitfensters um t0 den Wert 0 hat (d.h. $f(t-t_0)=0$ wenn $| t-t_0 | \geq t_W$) und zwischen diesen Randwerten interpoliert. Das so erhaltene modifizierte Innovationssignal $Inno_{w,k}(t) = Inno(t) \cdot f(t-I_{target}(k))$ ist in Fig. 3 ebenfalls gezeigt. Das Maximum dieser Funktion wird bestimmt und durch Abzug einer kurzen Vorlaufzeit τ_{pre} der Auswahlpunkt I(k) berechnet:

$$I(k) = \arg\max(Inno_{w,k}(t)) - \tau_{pre}$$

[0058] Die Vorlaufzeit τ_{pre} wird abhängig von dem Fenstertyp typischer Weise mit einem Wert zwischen 0,1 und 1 s gewählt. Dieses Verfahren ergibt einen Gesamtkompressionsfaktor C_{tot}, der den gewünschten Wert gut annähert.

[0059] Es ist auch möglich, das Maximum des unmodifizierten Innovationssignals Inno(t) im Fenster um $t_0=I_{target}(k)$ zu suchen. Dies entspricht der Verwendung einer Fensterfunktion, die 1 innerhalb des Zeitfensters ($|t-t_0| < t_W$) ist, jedoch 0 sonst.

[0060] Wenn diese Verfahren keine Gesamtkompression ergeben sollten, die dem gewünschten Wert für C_{tot} ausreichend nahe kommen, können die Beginnzeiten O(k) zum Kompensieren dieser Abweichung angepasst werden:

[0061] $O(k) = I(k) / C_{tot}$.

[0062] Im Falle einer zeitabhängigen Definition von $C_{tot}(t)$ wird die Anpassung der Beginnzeiten O(k) berechnet nach:

[0063] $O(k) = [I(k) - I(k-k_1)]/ C_{tot}(t) + O(k-k_1)$.

ANALYSEMODUL - ERZEUGEN DES INNOVATIONSSIGNALS

[0064] Das Innovationssignal Inno(t) kann zeitdiskret, wie z.B. eine Sequenz von aus Metadaten erzeugten Markern, oder kontinuierlich sein. Während bestimmte bekannte Verfahren ein als Innovationssignal geeignetes Signal erzeugen können, wie z.B. eine „gleitende" Mittelung der Signalleistung, ergaben sich die folgenden Verfahren als besonders zweckmäßig:

[0065] Eine erste Vorgehensweise geht von dem digitalisierten Klangsignal s1(n) aus - hierbei ist n der diskrete Zeit-Index-, um eine nichtlineare Größe y(n) zu berechnen:

$$y(n) = s1(n)^2\text{-}s1(n-1)s1(n+1);$$

sodann wird ein zeitliche Mittelung dieser Größe als Innovationssignal verwendet,

$$Inno(n) = A(n) = A\nu(y(n)).$$

[0066] Die Mittelung $A\nu$ erfolgt dadurch, dass der gleitende Mittelwert in einem Zeitintervall konstanter Länge um die aktuelle Zeit genommen wird, oder durch exponentielles Glätten; typische Zeitkonstanten liegen im Bereich von 0,3 bis 1 s. Dieses Verfahren ist effizient, benötigt nur geringen Rechenaufwand und betont hochfrequente Komponenten, die typisch für transiente Vorgänge sind. Weiters approximiert dieses Verfahren die frequenzabhängige Empfindlichkeit des menschlichen Gehörs.

[0067] Eine stärker differenzierte Vorgehensweise nützt auch die Zeitableitung der gemittelten Größe A(n),

$$dA(n)/dn = A(n) - A(n-m),$$

mit einem geeigneten Wert für m, wie z.B. 0,05 bis 0,5 s. Diese Zeitableitung zeigt den Anstieg der Leistung an. Das Produkt

$$B(n) = A(n) \cdot dA(n)/dn$$

kann dann als Innovationssignal verwendet werden.

[0068] Eine andere Vorgehensweise beruht auf einer Teilung des Klangsignals in eine Zahl von Frequenzbändern, die über Verfahren wie DFT, Gammaton-Filter, Oktavfilter oder Wavelet-Transformation erhalten werden können. Für jedes Frequenzband j = 1,...J mit zugehörendem Bandsignal x_j wird eine gleitende Mittelung der Leistung bestimmt,

$$P_j(n) = A\nu(x_j(n)^2),$$

mit einer Mittelungszeit von 0,5 bis 3 s. Aus dem Satz von Leistungen $P_j(n)$, der als Vektor P(n) mit Dimension J behandelt wird, wird das Innovationssignal über die euklidische Distanz zwischen Vektoren in einem gegebenen Zeitabstand m von typischer Weise 0,1 bis 1 s berechnet,

$$Inno(n) = \| P(n) - P(n-m) \|$$

worin $\| ... \|$ die üblichen euklidische Norm eines J-dimensionalen Vektors bezeichnet.

[0069] Das Gammaton-Filter ist ein Hörsignalfilter, das von R.D. Patterson entworfen wurde. Das Gammaton-Filter ist dafür bekannt, dass es den Respons der Basilarmembran gut simuliert. Siehe: Moore, B. und Glasberg, B. (1983). 'Suggested formulae for calculating auditory filter bandwidths and excitation patterns' („Formelvorschläge zum Berechnen von Hörsignalfilter-Bandbreiten und Erregungsmustern"), J. of the Acoustical Society of America, 74:750-753.

[0070] Noch eine andere Vorgehensweise setzt Clustering von Signal-Merkmalsvektoren ein. Das Klangsignal wird in Blöcke gleicher Länge geteilt, typischerweise von 10 bis 30 ms. Für jeden Block wird ein Signalmerkmalsvektor berechnet, beispielsweise mel-Frequenz-Cepstrumkoeffizienten (MFCC), die Signalleistung von Frequenzbändern, die Nulldurchgangsrate oder eine geeignete Kombination davon. Die Blöcke werden in „Meta-Blöcke" von vorzugsweise 20-100 aufeinanderfolgenden Blöcken gruppiert, entsprechend einer Länge von 0,2 bis 3 s. Die Zahl der Meta-Blöcke ist L. Für jeden Meta-Block werden aus den Signalmerkmalsvektoren der Blöcke in dem Meta-Block Parameter der Zentrumstendenz und optional Dispersionsparameter berechnet. Die so erhaltenen Parameter werden als „Meta-Merkmal" bezeichnet; der Satz von Parametern für jeden Meta-Block ergibt einen „Meta-Merkmal-Vektor". Die Werte jedes Meta-Merkmals, das über die L Meta-Blöcke vorkommt, werden dadurch standardisiert, dass der Mittelwerts des jeweiligen Meta-Merkmals über die L MetaBlöcke abgezogen und durch die Standardabweichung dividiert wird. Der standardisierte Meta-Merkmal-Vektor des 1-ten Metablocks (l = 1,...,L) wird im Folgenden als F(l) bezeichnet. Die Vektoren F(l) werden einem k-Means-Clustering-Verfahren mit einer typischen Clusterzahl k = 3 bis 30 unterworfen. Verfahren des k-Means-Qustering sind wohlbekannt und beruhen auf dem Konzept, Vektoren in Cluster aufzuteilen, sodass die gesamte Varianz der Vektordaten innerhalb eines Clusters minimiert wird. Das Ergebnis einer Clusteranalyse ist eine Gruppe von k Clustern mit wechselnder Zahl von Vektoren - in diesem Fall von Meta-Merkmal-Vektoren. Im einfachsten Fall findet ein Clustering-Durchlauf einmal für einen vorgegebenen Wert für k statt (Single-Level = einfache Runde; Multilevel-Clustering siehe unten). Ein Markersignal Mark(l) wird gemäß

$$Mark(l) = k^{-p} \text{ wenn } F(l) \text{ und } F(l-1) \text{ in verschiedenen Clustern liegen,}$$

$$0 \text{ sonst,}$$

erzeugt, wobei der Exponent p ein externer Parameter ist; günstige Werte sind p = 0,8 bis 3. (Der Wert k^{-p} ist beliebig für eine Einzel-Level, stellt jedoch einen Gewichtsfaktor bei dem weiter unten erläuterten Multilevel-Clustering dar.) Das Innovationssignal wird in Form des gemittelten Markersignals erhalten,

$$Inno(l) = Av(Mark(l)).$$

In diesem Fall ist exponentielles Glätten eine besonders günstige Art der Mittelung, mit einem Glättungsparameter a = 0,2 - 0.8, der rekursiv definiert werden kann gemäß:

$$Av(Mark(l)) = a \cdot Av(Mark(l-1)) + (1-a) \cdot Mark(l)$$

[0071] Vorzugsweise werden mehrere Clustering-Durchläufe („Levels" = „Runden") an den Meta-Merkmal-Vektoren eines Klangsignals durchgeführt, jeder Durchlauf mit einem verschiedenen Wert für die Clusteranzahl k. Mit anderen Worten, es wird eine Menge kg, g = 1,...,G, vorgegeben, und für jeden Wert k_g wird eine k-Means-Clusteranalyse durchgeführt. Die G Clusterergebnisse, die so erhalten werden, werden Levels genannt - daher der Name Multilevel-k-Means-Clustering. Das Markersignal $Mark_g(l)$ wird bei jeder Runde wie oben beschrieben ermittelt, und das Innovationssignal ist die gemittelte Summe der Markersignale,

$$Inno(l) = Av(\sum_g Mark_g(l)).$$

[0072] Eine nützliche Eigenschaft des Clustering-Verfahrens liegt darin, dass es schon dann gestartet werden kann, wenn nicht alle Datenvektoren vorhanden sind. Vielmehr können zusätzliche Datenvektoren zu einer Clusteranalyse hinzugefügt werden, die bereits angelaufen ist

oder sogar (vorläufig) konvergiert hat.

[0073] Eine andere Möglichkeit eines Innovationssignals ist ein „Novelty-Signal" („Neuigkeitssignal"), das von L. Lu, L. Wenyin, H. Zhang, in: 'Audio Textures: Theory and Applications' („Audiotexturen: Theorie und Anwendungen") - IEEE Trans. Speech and Audio Processing, Vol. 12, Nr. 2, März 2004, S. 156-167 behandelt wird. Das Novelty-Signal kann von Signalmerkmalen oder Meta-Merkmal-Vektoren abgeleitet werden.

GRAPHISCHE DARSTELLUNG VON AUDIOMATERIAL

[0074] Das Analysesignal 5, insbesondere das Innovationssignal Inno(t), bietet einen Weg zum Erzeugen einer graphischen Darstellung eines Audiosignals. Mittels einer solchen graphischen Darstellung können Blocke ähnlichen Inhalts ohne Umstände und viel leichter erkannt werden als in z.B. einem Spektrogramm (Diagramm der Energie über Zeit und Frequenz) oder einer Darstellung des Tonpegels (Lautstärke). Das nachfolgende Verfahren ist eine Erweiterung des Verfahrens, das von B. Logan and A. Salomon, in: 'A Music Similarity Function Based on Signal Analysis' („Eine auf Signalanalyse beruhende Musik-Ähnlichkeitsfunktion") - Proc. IEEE Int. Conf. On Multimedia and Expo (ICME'01), Tokyo 2001, vorgeschlagen wurde; diese Erweiterung wird in Kombination mit dem oben erläuterten Multilevel-k-Means-Clustering verwendet.

[0075] Fig. 4 zeigt ein Beispiel einer auf einem Innovationssignal basierenden graphischen Darstellung 40 eines Signals s1(t). Die gezeigte Darstellung gehört zu einem Drei-Level-k-Means-Clustering mit k_1=3, k_2=7 und k_3=15. Jedes Level entspricht jeweils einem (horizontalen) Streifen P1, P2, P3. Die Streifen zeigen Abfolgen von Mustern oder Farben, die je einen Cluster der jeweiligen Clusteranalyse repräsentieren. Intervalle, die zum selben Cluster gehören, sind mit jener Muster/Farbart markiert, die den Cluster identifiziert; jedes Mal, wenn der Meta-Vektor zu einem anderen Cluster wechselt, kann dieser Wechsel zusätzlich durch eine (vertikale) Trennlinie markiert sein.

[0076] Die Muster oder Farben können den Clustern beliebig zugeordnet sein, beispielsweise unter Verwendung von untereinander gut unterscheidbaren Muster/Farben. Als Alternative kann das Muster bzw. die Farbe durch einen Meta-Merkmal-Vektor bestimmt werden, der die Cluster repräsentiert (und z.B. als Zentroid der Meta-Merkmal-Vektoren F(l) des Clusters berechnet wurde). Beispielsweise können die Cluster-Meta-Merkmal-Vektoren in den Farbraum (in einer geeigneten Repräsentation wie RGB- oder CIE-Normvalenz-Farbenraum mit fester Luminanz) durch geeignete Reduktion der Dimension auf drei bzw. zwei Dimensionen mittels Hauptkomponentenanalyse abgebildet werden.

[0077] Die Wahl günstiger Werte k_g für die graphische Darstellung hängt auch von dem Kompressionsfaktor ab. So kann z.B. bei kleiner Kompression eine Kombination von Farbstreifen mit k_g= 7,15 und 30 einen guten Überblick ergeben, während bei einer hohen Kompression k_g= 2,4 und 7 geeignet sein kann. Fig. 4 zeigt einen Fall in der Mitte mit k_g= 3, 7 und 15.

ANWENDUNGSBEISPIELE

[0078] a) Suchmaschinen und Browserdienste

[0079] Das Internet ist zu einem wichtigen, wenn nicht dem hauptsächlichen, Verteilungsweg von Musik und anderen AVM geworden. Die Zahl der über Internet erreichbaren Lieferanten, Archiven und Privatsammlungen nimmt immer weiter schnell zu. Es ist absehbar, dass nur eine kleine Zahl dieser AVM geeignete Metadaten trägt, die einen ordentlichen Eindruck des jeweiligen Inhalts geben. Die Erfindung bietet einen Weg, eine für eine Schnellsuche geeignete Bestandsaufnahme zu gewinnen, um schneller durch diese Bestände navigieren zu können.

[0080] b) Überwachung

[0081] Die Sicherheitsdebatte nicht erst seit 9/11 hat zu einer starken Zunahme an Überwachungsaktivitäten im öffentlichen, privaten und geschäftlichen Bereich geführt. Die Untersuchung des aufgezeichneten Überwachungsmaterials nach auffälligen Ereignissen ist - naturgemäß und im Gegensatz zu Video - eine zeitaufwendige Aufgabe. Die Erfindung liefert einen

effektiven Zugang zu Erzeugen einer Übersicht von großen AVM-Mengen in kurzer Zeit.

[0082] c) Integrierte Metadaten-Editoren

[0083] Wie bereits erwähnt haben die europäischen Archive gewaltige Mengen von nicht annotiertem Audiovideomaterial. Um einen systematischen Zugriff und Überblick dieser AVM zu gestatten, müssen diese mit zeitsynchronen Metadaten versehen werden. Versuche, diesen Vorgang zu automatisieren, haben sich als schwierig herausgestellt und lieferten Fehler, die von Hand korrigiert werden mussten. Zum Zwecke der Korrektur und Kontrolle muss der Benutzer sich einen Überblick über das vorliegende AVM beschaffen. Die Erfindung erlaubt die Erzeugung eines solchen Überblicks auf schnellem Wege und auf Anfrage. Der Herstellungsaufwand der Annotierung von AVM kann somit deutlich verringert werden.

[0084] Die Genauigkeit der Darstellung kann abhängig von dem Fokuspunkt des Benutzers eingestellt werden. Der Benutzer wählt einen Zeitpunkt des AVM als Fokus und markiert dadurch diesen als „Gegenwart", die ungeändert (unkomprimiert) in Echtzeit wiedergegeben wird. Die Teile, die in der „Vergangenheit" oder „Zukunft" zu diesem Fokus liegen, werden komprimiert, mit einer mit zunehmendem Zeitabstand vom Fokus zunehmenden Kompression. Beispielsweise kann ein Zeitintervall bei 5 bis 4 min vor der Gegenwart auf 10 s kompaktiert werden, während ein Intervall zwischen 15 und 18 min gegenüber der Gegenwart auf 7 s zusammengezogen wird. Durch diese nichtlineare Kompression, die einer graphischen Zoom-Out-Funktion ähnlich ist, kann der Benutzer einen groben Überblick über die Inhalte außerhalb des Fokus erhalten, der gerade mit dem vorliegenden AVM verknüpft ist.

[0085] Im Rahmen der oben erwähnten fokusabhängigen Kompression kann eine Tonhöhenverschiebung den Zeitabstand von dem Fokus (der „Gegenwart") anzeigen. Somit hätte die entfernte „Vergangenheit" oder „Zukunft" eine höhere Tonlage als zur Gegenwart vergleichsweise nahe Teile, nicht unähnlich einer Schnellwiedergabe einer Bandaufnahme.

[0086] d) Akustische Thumbnails

[0087] Die Erfindung bietet auch einen einfachen Weg, Kurzdarstellung zu erzeugen, die als akustische „Fingerabdrücke" oder „Thumbnails" verwendbar sind. Diese akustischen Fingerabdrücke bieten einen intuitiven Zugang zu den dahinter steckenden AVM-Dateien, da das erfindungsgemäße Verfahren ein Zeitintervall auf eine Weise reduziert, das den grundlegenden kategoriellen Duktus des AVM beibehält, jedoch Details geringer Wichtigkeit unterdrückt. Ein solcher akustischer Thumbnail braucht nur eine kurze Zeit zum Laden oder Übertragen und könnte - wie die sogenannten Thumbnail-Ikons in Bildverzeichnissen - als ein „Earcon" oder „Ohr-kon" verwendet werden, was das Abfragen von zeitsparender Vorabinformation ermöglicht. Diese Ohrkons können getrennt erzeugt und verteilt oder verkauft werden, möglicherweise als Web-Dienst. Sie können auch als persönliche Klingeltöne in Mobiltelefonen oder ähnlichen Anwendungen verwendet werden.

[0088] Während in dieser Offenbarung bevorzugte Ausführungsformen der Erfindung gezeigt und beschrieben werden, versteht es sich, dass diese Ausführungsformen nur auf beispielhaftem Wege dargebracht sind. Zahlreiche Abwandlungen, Änderungen und Ersetzungen ergeben sich für den Fachmann, ohne von der Erfindung abzuweichen. Dem entsprechend ist es beabsichtigt, dass die beigefügten Ansprüche alle derartigen Abwandlungen abdecken, die in den Bereich und Sinn der Erfindung fallen.

Patentansprüche

1. Verfahren zum Bearbeiten von in einer Aufnahme enthaltenen Audio-Daten zum Gewinnen einer zum Anhören wiedergebbaren gekürzten Version, enthaltend die Schritte:
   - Auswahl einer Anzahl von aufeinander folgenden, überlappungsfreien Segmenten der Audiodaten;
   - Reduktion jedes Segments durch zeitliche Kompression; und
   - Kombinieren der so reduzierten Segmente.

   dadurch gekennzeichnet, dass die Auswahl von Segmenten der Audiodaten aufweist:

   Ableiten eines Innovationssignals aus den Audiodaten, wobei das Innovationssignal eine Größe darstellt, die eine Änderungsrate des Inhalts in den Audiodaten angibt;

   Bestimmen von Zeitpunkten von Maxima des Innovationssignals;

   Auswahl von Segmenten, die jeweils diese Zeitpunkte enthalten;

   Reduktion dieser Zeitpunkte durch jeweilige Zeitversetzungen; und

   Setzen von Segmentbeginnzeiten an den so reduzierten Zeitpunkten.

2. Verfahren nach Anspruch 1, bei welchem die zeitliche Kompression mit einem über die Zeit variierenden Kompressionsfaktor stattfindet, der zwischen den Segmenten variiert.

3. Verfahren nach Anspruch 1 oder 2, bei welchem die Berechnung des Innovationssignals ausgehend von einem Audiodaten-Signal s1(n) aufweist:

   Ableiten einer nicht-linearen Größe $y(n) = s1(n)^2 - s1(n-1) \cdot s1(n+1)$;

   Mittelung dieser nicht linearen Größe mit einer Glättungsfunktion A_ν, wobei sich eine gemittelte Größe $A(n) = A_\nu[y(n)]$ ergibt; und

   Verwendung dieser gemittelten Größe als Innovationssignal Inno(n).

4. Verfahren nach Anspruch 1 oder 2, bei welchem die Berechnung des Innovationssignals ausgehend von einem Audiodaten-Signal s1(n) aufweist:

   Ableiten einer nicht-linearen Größe $y(n) = s1(n)^2 - s1(n-1) \cdot s1(n+1)$;

   Mittelung dieser nicht linearen Größe mit einer Glättungsfunktion A_ν, wobei sich eine gemittelte Größe $A(n) = A_\nu[y(n)]$ ergibt; und

   Kombinieren dieser gemittelten Größe mit seinen vorangehenden Werten A(n-m) zur Berechnung eines Innovationssignals $Inno(n) = A(n)^2 - A(n) \cdot A(n-m)$.

5. Verfahren nach einem der Ansprüche 1 bis 4, bei welchem die Berechnung des Innovationssignals aufweist:

   Aufteilen eines Audiodaten-Signals in eine Anzahl von Frequenzband-Signalen;

   Bandpass-Filtern der Frequenzband-Signale;

   Berechnen eines wandernden Mittelwerts einer momentanen Leistung der so gefilterten Signale unter Verwendung einer Glättungsfunktion A_ν;

   Kombinieren der so erhaltenen Signale in einen mehrdimensionalen Leistungsvektor P(n); und

   Berechnen einer Distanzfunktion zwischen dem aktuellen und einem vorangehenden Wert des Leistungsvektors zum Bilden des Innovationssignals Inno(n) = dist[P(n) - P(n-m)].

6. Verfahren nach einem der Ansprüche 1 bis 5, bei welchem die Berechnung des Innovationssignals aufweist:

   Aufteilen eines Audiodaten-Signals in eine Anzahl von Frequenzband-Signalen;

   Berechnen einer entsprechenden Zahl sekundärer Signale aus den Frequenzband-Signalen mithilfe zumindest eines der folgenden Verfahren: Filtern des Signals, Glätten des Signals, und/oder Berechnen eines lokalen Polynoms aus dem Signal;

   Kombinieren der sekundären Signale in einen mehrdimensionalen Leistungsvektor P(n); und

   Berechnen einer Distanzfunktion zwischen dem aktuellen und einem vorangehenden Wert des Leistungsvektors zum Bilden des Innovationssignals Inno(n) = dist[P(n) - P(n-m)].

7. Verfahren nach Anspruch 1 bis 6, bei welchem die Berechnung des Innovationssignals aufweist:

Segmentieren der Audiodaten in überlappungsfreie Segmente;

Berechnen eines Meta-Merkmal-Vektors F(l) für jedes dieser Segmente;

Durchführen einer k-Means-Clusteranalyse für die so erhaltenen Meta-Merkmal-Vektoren; und

Berechnen eines Markersignals für jedes Segment zum Erhalt des Innovationssignals durch Zuweisen eines positiven Werts dann, wenn der Meta-Merkmal-Vektor in einem von dem Cluster des vorangehenden Segments verschiedenen Cluster liegt, bzw. eines Wertes Null ansonsten.

8. Verfahren nach Anspruch 7, bei welchem die k-Means-Clusteranalyse für G verschiedene Werte der Zahl k_g der Cluster mit g=1,.. .,G vorgenommen wird, wobei G Markersignale für jedes Segment erhalten werden, und das Innovationssignal durch Mitteln einer Überlagerung dieser Markersignale unter Verwendung einer Glättungsfunktion Av zum Erhalt des Innovationssignals $Inno(l) = Av(\sum_g Mark_g(l))$ berechnet wird.

9. Verfahren nach Anspruch 8, bei welchen die Berechnung der G Markersignale gemäß $Markg(l) = h(k_g)$, wenn F(l) und F(l-1) in verschiedenen Clustern liegen, bzw.

0 sonst,

mit einer monoton fallenden Funktion h vorgenommen wird.

10. Verfahren nach einem der Ansprüche 7 bis 9, bei welchem die Berechnung der Meta-Merkmal-Vektoren ein Aufteilen der Segmente der Audiodaten in Untersegmente enthält,

Berechnen von Merkmalsvektoren für diese Untersegmente;

Berechnen von Verteilungsparametern dieser Merkmalsvektoren; und

Kombinieren dieser Verteilungsparameter in einen Meta-Merkmal-Vektor.

11. Verfahren nach einem der Ansprüche 1 bis 10, bei welchem der Schritt der Segmentierung der Audiodaten auf Nicht-Audio-Daten beruht, die in der Aufnahme enthalten und synchron mit den Audiodaten sind, wobei die Segmentbeginnzeiten bei in den Nicht-Audio-Daten vorhandenen Zeitmarkern platziert werden.

12. Verfahren nach einem der Ansprüche 1 bis 11, bei welchem der Schritt des Kombinierens der reduzierten Segmente in chronologischer Reihenfolge in Bezug auf ihre ursprüngliche Position in den Audiodaten vorgenommen wird, unter Auswahl entweder der voran oder der rückwärts laufenden Ordnung.

13. Verfahren nach einem der Ansprüche 1 bis 12, bei welchem der Schritt des Kombinierens der reduzierten Segmente eine Überlagerung der Segmente enthält.

14. Verfahren nach Anspruch 13, bei welchem die Überlagerung der Segment eine gestaffelte Überlagerung ist/enthält, wobei die Segmente zu aufeinander folgenden Beginnzeiten anfangen und jedes nicht-erste Segment eine Beginnzeit innerhalb der Dauer des jeweils vorangehenden Segments hat.

15. Verfahren zum Bearbeiten von Audio-Daten zum Gewinnen einer graphisch darstellbaren Version, enthaltend die Schritte:

Ableiten eines Innovationssignals aus den Audiodaten, wobei das Innovationssignal eine Größe darstellt, die eine Änderungsrate des Inhalts in den Audiodaten angibt;

Bestimmen von Zeitpunkten von Maxima des Innovationssignals;

Setzen von Segmentgrenzen an so bestimmten Zeitpunkten; und

Anzeigen der so definierten Segmente in einer linearen Abfolge von Flächen mit variierender graphischer Wiedergabe.

Hierzu 2 Blatt Zeichnungen

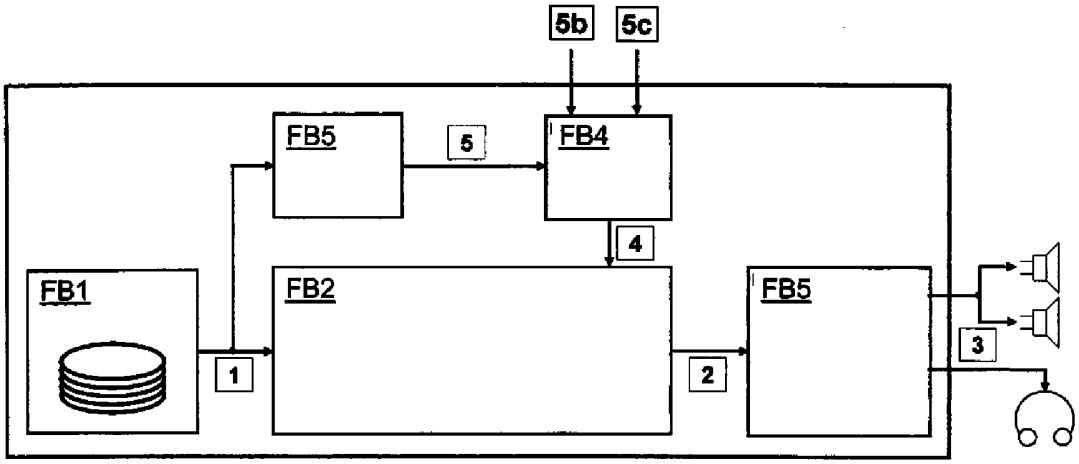
5b
5c
FB5
5
FB4
4
FB1
FB2
FB5
1
2
3

Fig. 1

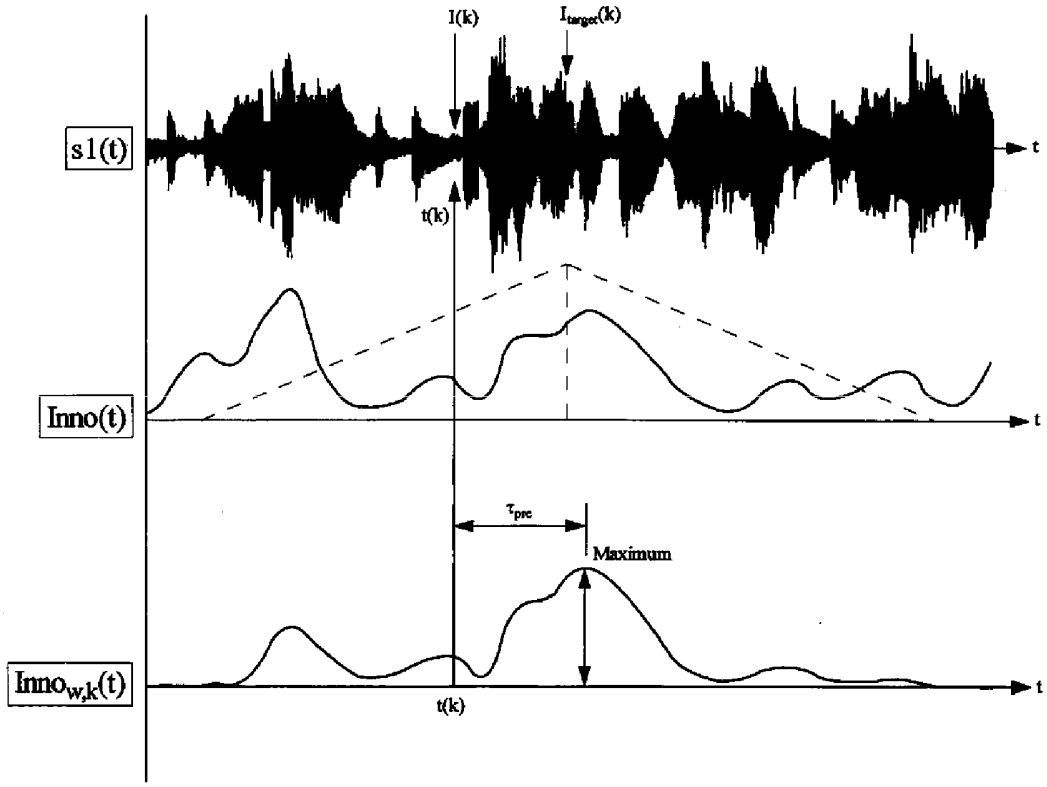
I(k)
$I_{target}(k)$
s1(t)
t(k)
t
Inno(t)
t
τ_{pre}
Maximum
$Inno_{w,k}(t)$
t
t(k)

Fig. 3

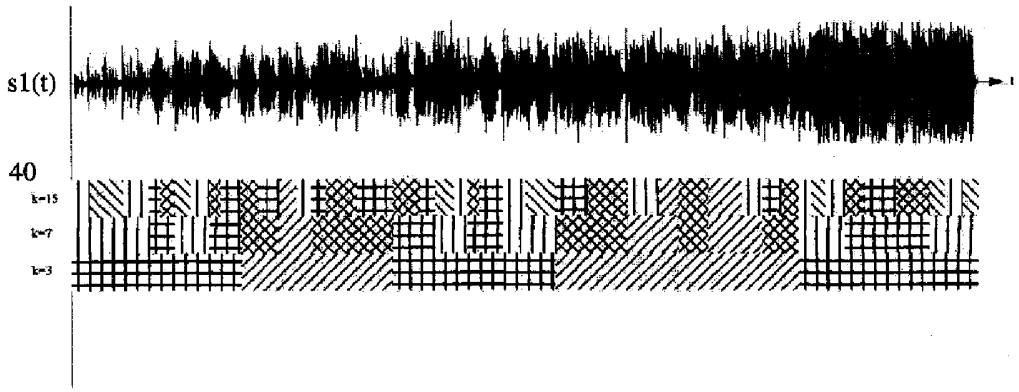
s1(t)
t
40
k=15
k=7
k=3

Fig. 4

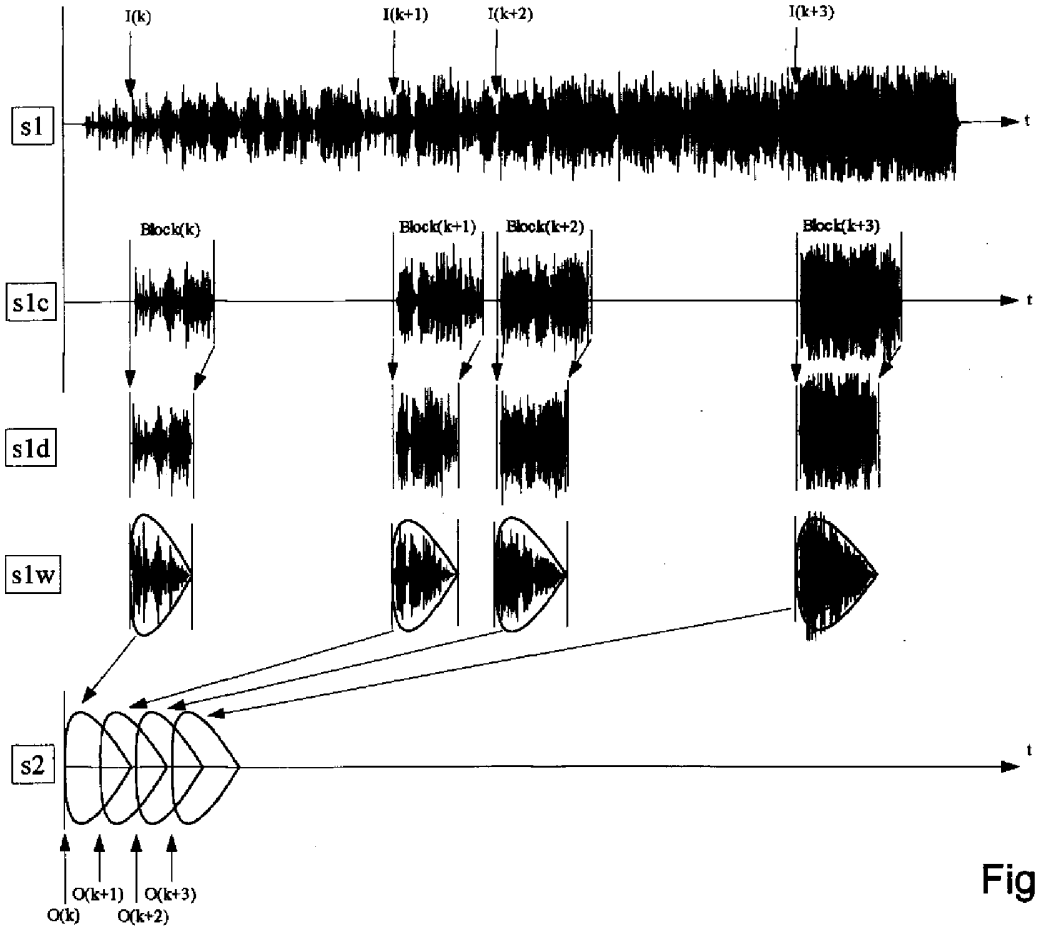
I(k)
I(k+1)
I(k+2)
I(k+3)
s1
t
Block(k)
Block(k+1)
Block(k+2)
Block(k+3)
s1c
s1d
s1w
s2
O(k)
O(k+1)
O(k+2)
O(k+3)

Fig. 2