Multimedia-Datenbanken

Kapitel 5: Multimedia-Daten – Audio

Friedrich-Alexander-Universität Erlangen-Nürnberg
Technische Fakultät, Institut für Informatik
Lehrstuhl für Informatik 6 (Datenbanksysteme)

Prof. Dr. Klaus Meyer-Wegener

Wintersemester 2002 / 2003

Technische Universität Kaiserslautern
Fachbereich Informatik
AG Datenbanken und Informationssysteme

Dr. Ulrich Marder

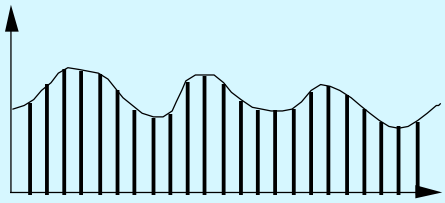
Wintersemester 2003 / 2004

5.1 Tonaufnahme (Audio)

- **meistens Sprache oder Musik – aber nicht nur!**
- **Rohdaten:**
  - Folge von Energieniveaus (Lautstärkepegeln) oder Frequenzanteilen (Fourier-Analyse eines Zeitfensters)
  - wegen des enormen Datenvolumens immer komprimiert
- **Abtasttheorem (Nyquist-Theorem):**
  - Abtastung des Energieniveaus muss mindestens doppelt so häufig erfolgen wie höchste zu erfassende Frequenz
  - Telefon: 3000 Hz
    MW-Radio: 4000 Hz
    UKW-Radio: 8000 Hz
    Hifi: 22000 Hz
- **Beispiel Audio-CD:**
  - 44100 Messwerte pro Sekunde und pro Stereokanal,16 bit pro Messwert: 176,4 KB pro Sekunde, ca. 10 MB pro Minute, 635 MB pro Stunde

Tonaufnahme (2)

- **Registrierungsdaten:**
  - **Auflösung** (Resolution):
    Anzahl unterschiedener Energieniveaus
    oft 256, d. h. 8 bit pro Messwert
  - **Aufzeichnungsfrequenz** (Sampling Rate)
  - Anzahl der Kanäle (1 bei Mono, 2 bei Stereo)
- **Wellenform-Codierung (Waveform Encoding)**
  - ausgehend vom Sampling, also Amplitudenmessung in festem Zeitabstand



  - auch als "Pulse Code Modulation" (PCM) bezeichnet,
    weil Wellenform als getaktetem Impuls aufmoduliert erscheint

Tonaufnahme (3)

- **Logarithmisches PCM**
  - Schrittweite des Quantisierens nicht konstant,
    sondern bei niedrigen Werten kleiner als bei hohen
    - Rauschverminderung bei leisen Passagen
    - z.B. μ-LAW (Telefon Nordamerika und Japan),
      A-LAW (Telefon Europa, Rest der Welt und internationale Leitungen)
  - weniger Bits reichen aus, um gleiche Amplitude zu überdecken
    - grob: μ-LAW mit 8 Bit entspricht linearer Quantisierung mit 12 Bit,
      μ -LAW mit 12 Bit linearer mit 16 Bit
- **Differenzen-PCM (DPCM)**
  - aufeinanderfolgende Abtastwerte fast immer eng korreliert
  - für jeden Wert zunächst Prädiktor berechnen,
    dann nur Differenz zwischen ihm und tatsächlichem Wert speichern
  - $p(x_i) = a_1 x_{i-1} + a_2 x_{i-2} + \ldots$
  - häufig einfach: $p(x_i) = x_{i-1}$

Tonaufnahme (4)

- **Differenzen-PCM (Forts.)**
  - bei 256 Quantisierungsstufen selten mehr als 32 Stufen Differenz zwischen aufeinanderfolgenden Messwerten
    → 5 Bit statt 8

unkomprimiert	112	114	117	115	111	109
Differenzen		+2	+3	-2	-4	-2

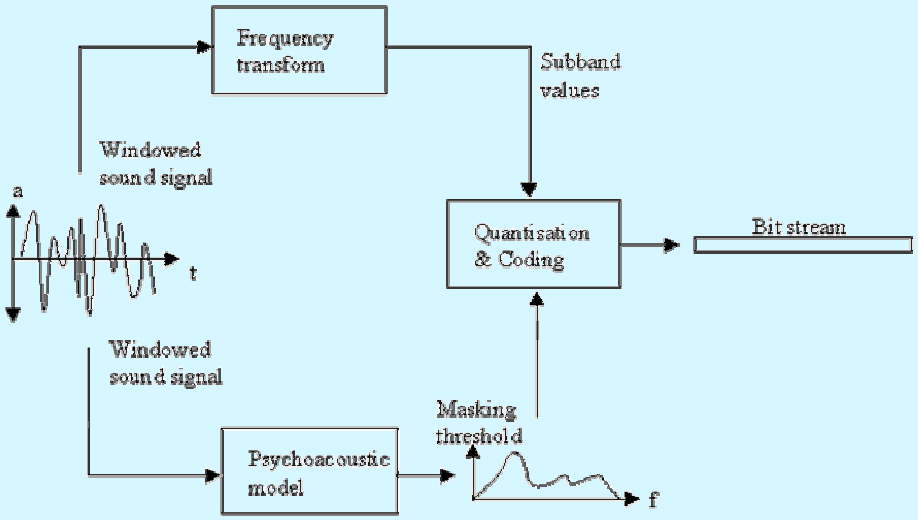
- **Delta-Modulation (DM)**
  - Anzahl der Quantisierungsstufen, die für Differenzen benötigt werden, um so geringer, je kürzer das Abtastintervall
  - → so klein machen, dass nur noch 1 Bit nötig
  - einfach!
    bei niedrigen Bitraten (32 kbit/s, Telefonqualität) besser als alle anderen Verfahren

Tonaufnahme (5)

- **weitere Verbesserung: Adaptives DPCM (ADPCM)**
  - Berechnung Prädiktor aus mehreren vorhergehenden Abtastwerten
  - adaptiv: Änderung der Auflösung
    - gering bei starken Schwankungen (laut), hoch bei schwachen (leise)
- **Parameter-Codierung (nur Sprache)**
  - Modell des menschlichen Sprechapparats:
    - Schwingungsfrequenz der Stimmbänder, Formung von Mund- und Rachenraum, ausgeblasene Luft usw.
    - durch Parameter beschreiben
  - Ermittlung durch Spektralanalyse eines Abschnitts
    - Fenster, im Millisekundenbereich
  - z. B. Linear Predictive Coding (LPC)
    - beim nächsten Fenster nur Differenz zum vorhergehenden speichern ("Frequenz x um y stärker")
- **(beide, Wellenform- und Parameter-Codierung durch Spezial-Hardware realisiert)**

Tonaufnahme (6)

- **MPEG-1 Audio**
  - Layers I, II und III (MP3)
  - Bitraten [kbps]
    - 384 (1:4)
    - 256-192
    - 128-112 (CD-Qualität)
    - bis 8 bei MPEG-2.5
  - Psychoakustisches Modell
  - Redundanz des zweiten Kanals
  - Variable Bitrate (nur Layer III)



Tonaufnahme (7)

- **Beschreibungsdaten:**
  - bei Sprache Transkription als Text
  - bei Musik Notenschrift oder MIDI (s. unten)
  - Strukturinformation: Pausen (wichtig für Editieren)
- **Operationen:**
  - Eingabe
    - von Datei in einem bestimmten Format oder direkt vom Gerät (Echtzeit!)
  - Ausgabe
    - auf Datei oder direkt auf Gerät (Echtzeit)
  - Modifikation
    - "Schneiden" wie in einem Tonstudio; Positionsangaben entweder Zeit oder laufende Nummer des Messwerts
    - "Verkleben" entsprechend
    - Lautstärke erhöhen (schwierig: nicht linear)

Tonaufnahme (8)

- **Operationen (Forts.)**
  - Auswertung, Aggregation
    - statistische Angaben über Verteilung der Messwerte
    - Auffinden von Pausen (Schwellenwert für Amplitude)
    - Worterkennung (bei Sprache)
  - Vergleich (Suche)
    - Pattern Matching
    - Gleichheit oft zu restriktiv, Ähnlichkeitsmaße?!
    - am ehesten wieder über Beschreibungsdaten, z. B. Text
- **Subtypen**
  - gesprochene Sprache (am wichtigsten)
  - Musik
  - Maschinengeräusch (Kfz-Motor)
  - u. v. a.

Musik: MIDI

- **„Music Instrument Digital Interface“**
  - seit 1983 von der Musikindustrie verwendet
  - Definition einer Schnittstelle zwischen elektronischen Musikinstrumenten (und auch zu Rechnern)
- **instrumentenbezogene Darstellung:**
  - Bezeichnung des Instruments, Beginn und Ende einer Note, Grundfrequenz, Lautstärke u. a.
  - 10 Oktaven, d. h. 128 Noten
- **für 10 Minuten Musik ca. 200 KB MIDI-Daten – wesentlich weniger als bei Sampling!**
  - Keyboard → Computer: Eingabe,
    Computer → Synthesizer: Ausgabe
  - Sequenzer: Gerät zum Zwischenspeichern und ggf. Verändern der Daten – kann Computer mit entspr. Software sein
- **MIDI-Standard („General MIDI“):**
  - 16 Kanäle mit jeweils einem Synthesizer-Instrument,
    128 Instrumente (z. B. 0 = "Acoustic Grand Piano"),
    3–16 Noten pro Kanal

5.2 Indexierung und Retrieval von Audio

- **einfachste Methode: über Titel und Dateiname ...**
  - sehr verbreitet
  - Namen allerdings unvollständig und subjektiv – schwierig zu finden
  - außerdem keine Möglichkeit, Audio-Aufnahmen zu finden, die so klingen wie etwas, was gerade zu hören ist
- **Inhalt verwenden**
  - Vergleich Messwert für Messwert
    - wenig erfolgversprechend, da Unterschiede in Abtastrate und Auflösung nicht berücksichtigt
  - daher Merkmale (Features) extrahieren und nutzen
    - mittlere Amplitude
    - Frequenz-Verteilung

Allgemeinerer Ansatz für Audio-Retrieval

- **Klassifikation**
  - in verbreitete Typen wie Sprache, Musik, Geräusch
- **differenzierte Behandlung jeder Klasse**
  - z. B. Sprache: Spracherkennung und Indexierung des Textes
- **Anfragen**
  - ebenso klassifiziert, verarbeitet und indexiert
- **Retrieval**
  - beruht auf der Ähnlichkeit der Anfrage-Merkmale mit den Merkmalen der gespeicherten Tondokumente

Klassifikation

- **wichtig aus einer Reihe von Gründen:**
  1. verschiedene Typen verlangen unterschiedliche Verarbeitung und unterschiedliche Indexierungstechniken
  2. verschiedene Typen haben unterschiedliche Bedeutung für eine Anwendung
  3. Sprache ist der wichtigste Typ, und es gibt heute recht erfolgreiche Spracherkennungs-Techniken und –Systeme
  4. die Typinformation selbst ist in einigen Anwendungen sehr nützlich
  5. der Suchraum reduziert sich auf eine Klasse

5.2.1 Audio-Eigenschaften und -Merkmale

- **Basis für Klassifikation und Retrieval**
- **zwei Darstellungen**
  - **Zeit-Domäne** (Amplitude über der Zeit)
  - **Frequenz-Domäne** (Stärke über Frequenz)
- **jeweils unterschiedliche Merkmale**
- **zusätzlich weitere Merkmale**
  - subjektiv
  - z. B. Timbre

Merkmale in der Zeit-Domäne

- **Amplitude**
  - Druckschwankung um den Normaldruck herum
  - Stille = Amplitude null
- **durchschnittliche Energie**
  - charakterisiert die Lautstärke des Audio-Signals

$$E = (\sum_{n=0}^{N-1} x(n)^2) / N$$

mit *E* als durchschnittlicher Energie,
N als Gesamtzahl aller Messwerte
und *x(n)* als Messwert Nr. *n*

Merkmale in der Zeit-Domäne (2)

- **Nulldurchlaufsrate (Zero-crossing rate)**
  - charakterisiert die Häufigkeit des Vorzeichenwechsels im Signal und in gewissen Maße auch die durchschnittliche Frequenz des Signals

$$ZC = \left(\sum_{n=1}^{N} | \operatorname{sgn} x(n) - \operatorname{sgn} x(n-1) | \right) / 2N$$

mit sgn *x(n)* als Vorzeichen von *x(n)*; 1 wenn *x(n)* positiv, –1 sonst

- **Anteil der Stille (silence ratio)**
  - Anteil der Messwerte an der Gesamtzahl, die einer Periode (!) der Stille angehören
  - zwei Schwellenwerte:
    - Amplitudenwert, unterhalb dessen Stille angenommen wird
    - Anzahl unmittelbar aufeinanderfolgender Messwerte, die mindestens still sein müssen, um eine Stilleperiode zu bilden

Merkmale in der Frequenz-Domäne

- **Fourier-Transformation des Signals**
  - Zerlegung in Frequenz-Anteile mit Faktoren (Koeffizienten)
  - Darstellung: Faktoren über Frequenz (Energiemenge pro Frequenz in Dezibel, dB)
  - auch **Spektrum** des Signals genannt
- **Bandbreite**
  - Intervall der vorkommenden Frequenzen
  - bei Musik größer als bei Sprache
  - Differenz von größter und kleinster Frequenz im Spektrum
  - nur Frequenzen mit Energiemenge größer 3 dB betrachtet

Merkmale in der Frequenz-Domäne (2)

- **Energieverteilung**
  - direkt aus dem Spektrum ablesbar
  - Frequenzen mit hoher Energie: nützlich bei der Klassifikation, Musik hat mehr Frequenzen mit hoher Energie als Sprache
  - Berechnung von **Frequenzbändern** mit hoher und niedriger Energie
    - Schwellenwert, z. B. 7 kHz
    - Energie pro Band: Summe der Energien aller Frequenzen im Band
  - Zentroid: Mittelpunkt der spektralen Energieverteilung
    - bei Sprache niedriger als bei Musik
    - auch Brightness genannt

Merkmale in der Frequenz-Domäne (3)

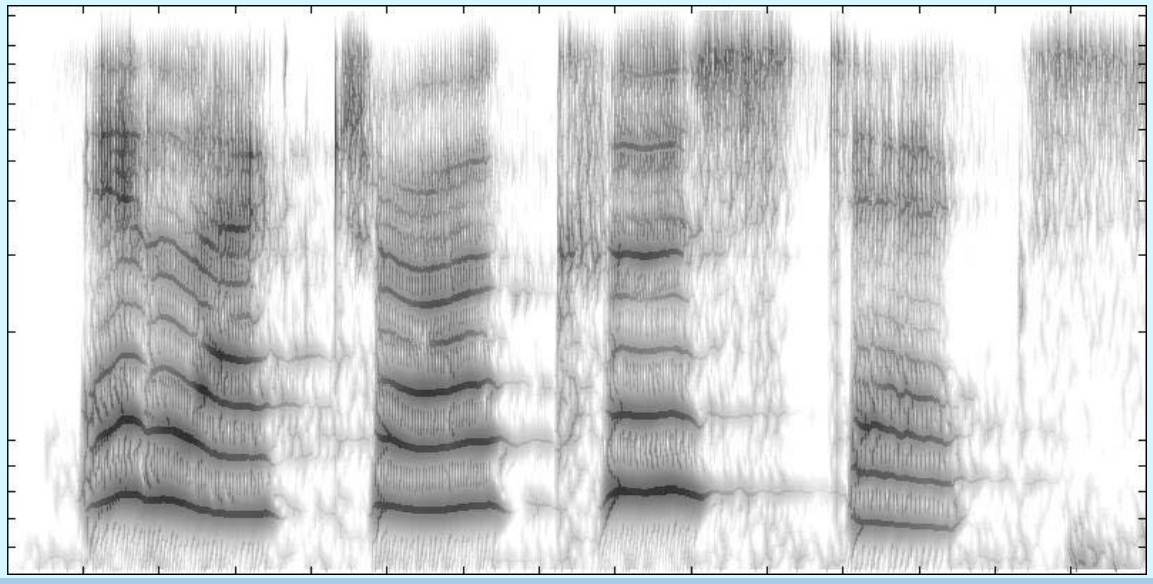
- **Harmonie**
  - spektrale Komponenten oft Vielfache der niedrigsten und lautesten Frequenz ("fundamental frequency")
  - Musik in der Regel harmonischer als andere Geräusche
  - Prüfung, ob eine Tonaufnahme harmonisch ist:
    dominante Komponenten Vielfache der fundamentalen Frequenz?
  - Beispiel: Flöte spielt Note G4;
    Spitzen bei den Frequenzen 400 Hz, 800 Hz, 1200 Hz, 1600 Hz usw.
  - f, 2f, 3f, 4f usw. Harmonische der Note
- **Pitch**
  - nur periodische Klänge (Instrumente, Stimme)
  - Perkussion dagegen nicht
  - subjektiv, verwandt, aber nicht gleichbedeutend mit der fundamentalen Frequenz die (oft als Näherung verwendet wird)

Spektrogramm

- **einfache Darstellungen haben Grenzen:**
  - Zeit-Domäne zeigt die Frequenz-Anteile eines Signals nicht
  - Frequenz-Domäne zeigt nicht, wann die Frequenzen auftreten
- **kombinierte Darstellung**
  - (Rasterbild, Matrix von Bildpunkten)
  - x-Achse: Zeit
  - y-Achse: Frequenzanteile
  - Schwärzung eines Punkts: Energie der Frequenz zu dieser Zeit
- **Analysen**
  - Regularität des Auftretens von Frequenzen
  - Musik ist regulärer als andere Geräusche

Spektrogramm (2)

- **Beispiel**
  - weibliche Sprecherin, (englisch "Electroacoustics"), Signaldauer 1,5 s
  - Quelle: http://www.mmk.ei.tum.de/~rue/mum/eurospeech99/demo/



5.2.2 Klassifikation

- **Nutzung der Merkmale**
- **hier nur Musik und Sprache**
  - könnten weiter differenziert werden:
    - Arten von Musik
    - männliche oder weibliche Sprache

Hauptcharakteristiken

- **Sprache**
  - Bandbreite vergleichsweise gering, 100 bis 7000 Hz
  - Zentroid deshalb niedriger als bei Musik
  - häufige Pausen (zwischen Worten und Sätzen) – höherer Anteil der Stille
  - charakteristische Struktur: Folgen von Silben, die aus kurzen Perioden von Friktionen (Konsonanten) bestehen, auf die längere Perioden von Vokalen folgen – während der Friktionen hohe Nulldurchlaufrate, ZC variiert stärker
- **Musik**
  - hohe Bandbreite, 16 bis 20.000 Hz
  - Zentroid deshalb höher
  - niedriger Anteil der Stille
    - Ausnahmen: Soloinstrument, A-Capella-Gesang
  - Nulldurchlauf variiert nicht so stark
  - regular beat

Klassifikationssysteme

- **Schritt für Schritt**
  - ein Merkmal nach dem anderen
  - z. B. erst Zentroid
    - wenn hoch: Musik

    dann Anteil der Stille
    - wenn niedrig: Musik

    dann ZC-Variabilität
    - wenn niedrig: Solo-Musik
    - sonst: Sprache
  - Reihenfolge wichtig
    - algorithmische Komplexität und Differenzierungsvermögen
    - einfach zu berechnen und hohe Differenzierung zuerst
  - ein Merkmal allein auch schon nutzbar:
    - nur ZC: bis zu 90 % korrekt klassifiziert
    - nur Anteil der Stille: bis zu 82 %

Klassifikationssysteme (2)

- **Feature-Vektoren**
  - Werte einer Menge von Merkmalen berechnet und zu Vektor zusammengefasst
  - Training: Durchschnittsvektor (Referenzvektor) einer jeden Klasse bestimmen
  - für neues Audio Feature-Vektor berechnen und Distanz zu den Referenzvektoren ermitteln (meist euklidisch)

5.2.3 Spracherkennung

- **nach der Klassifikation**
- **Techniken**
  - Time Warping (Sprechgeschwindigkeit)
  - Hidden Markov Models
  - neuronale Netze
- **Leistung**

Gegenstand	*Typ*	*Vokabular*	*Fehlerrate in %*
Ziffern	**gelesen**	**10**	**< 0,3**
Flugbuchungs-system	**spontan**	**2500**	**2**
Wall Street Journal	**gelesen**	**64000**	**7**
Radio-nachrichten	**gelesen / spontan**	**64000**	**30**
Telefonanruf	**spontan**	**10000**	**50**

5.2.4 Musik-Indexierung

- **noch in den Anfängen**
- **zwei Typen**
  - strukturierte / synthetische Musik (MIDI)
  - aufgezeichnete Musik
- **Indexierung strukturierter Musik**
  - keine Extraktion von Merkmalen erforderlich
  - sogar exakte Übereinstimmung als Suchmethode denkbar
  - allerdings mag das eingestellte Instrument ein anderes als das gewünschte sein
  - Ähnlichkeit schwierig zu definieren
  - eine Möglichkeit: nur den Pitch-Wechsel berücksichtigen
    - Up, Down, Repeat – U, D, R
    - Parsons, D., *The Directory of Tunes and Musical Themes*, Spencer Brown, 1975 (vergriffen)
    - Melodyhound: http://name-this-tune.com/ (Uni Karlsruhe)

  Retrieval dann: Zeichenkettenvergleich

Musik-Indexierung (2)

- **aufgezeichnete Musik (sample-based)**
  - Anfrage: Vorsingen oder -summen
  - Menge von Merkmalen
    - z. B. Lautstärke, Pitch, Brightness, Bandbreite und Harmonie
    - Vektoren und Distanzberechnung
  - Pitch
    - für jede Note extrahieren oder schätzen ("pitch tracking")
    - Segmentierung in Noten nicht einfach
      (am besten Pause zwischen den Noten ...)
    - Darstellung dann Folge von Pitches,
      ggf. auch wieder mit Up, Down und Similar
    - Ähnlichkeit: in der Folge dürfen k Pitches falsch sein