

Chapter 5 - Audio



Audio Recording

- Usually speech or music, possibly arbitrary sound
- Digitization: Pulse Code Modulation - PCM
 - sampling, i.e. amplitude measurements, quantization at fixed time intervals
- Sampling theorem (Nyquist-Shannon theorem)
 - fundamental result in information theory
 - sampling of energy levels has to occur with at least double the frequency of the highest frequency occurring in the signal
 - phone: 3000 Hz, AM-radio: 4000 Hz, FM-radio: 8000 Hz
Hifi: 22000 Hz (~ maximum frequency recognized by the human ear)
- Example Audio-CD:
 - 44100 sampling points per second and per stereo channel:
176,4 KB per second, approx. 10 MB per minute, 635 MB per hour
- Encoding/Compression
 - Waveform Encoding (based on PCM)
 - Parameter Encoding



Waveform Encoding

- Logarithmic PCM
 - quantization intervals are not constant, but smaller for lower amplitude values
 - noise reduction in softer passages
 - e.g., μ -LAW (phone in North America and Japan), A-LAW (phone in Europe, rest of the world and international phone lines)
 - fewer bits are sufficient to cover the same amplitude
 - μ -LAW with 8 bit roughly equivalent to linear quantization with 12 bit,
 - μ -LAW with 12 bit roughly equivalent to linear with 16 bit
- Differential PCM (DPCM)
 - subsequent sample values are often correlated
 - for each value, first compute a predicted value, then store the difference between the predicted and the actual value
 - $p(x_i) = a_1 x_{i-1} + a_2 x_{i-2} + \dots$
 - or simply: $p(x_i) = x_{i-1}$

Waveform Encoding (2)

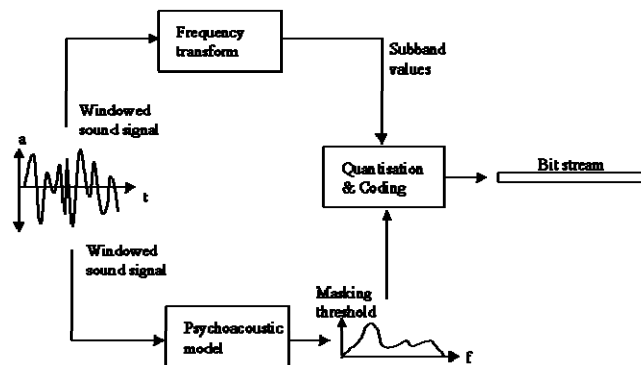
- Differential PCM (cont.)
 - with 256 quantization levels, differences between subsequent sample values of more than 32 levels rarely occur
→ 6 bits are sufficient

uncompressed	112	114	117	115	111	109
differences		+2	+3	-2	-4	-2

- Delta Modulation (DM)
 - number of quantization levels required for recording differences becomes smaller for shorter sampling intervals
 - → reduce sampling intervals until 1 bit is sufficient for recording the difference
 - for low bitrates (32 kbits/s, phone line quality) improvements over other approaches
- Adaptive DPCM (ADPCM) provides further improvement
 - prediction based on multiple preceding sample values
 - adaptive: resolution can vary
 - *high* for strong variations, *low* for weak variations

Waveform Encoding (3)

- MPEG-1 Audio
 - Layers I, II and III (MP3)
 - Bitrates [kbps]
 - 384 (1:4)
 - 256-192
 - 128-112 (CD quality)
 - psycho-acoustic model
 - redundancy of second channel
 - variable bitrate (Layer III only)



Parameter Encoding

- Parameter encoding (only for speech)
 - based on a model of the human vocal tract:
 - pitch of the voice based on oscillation frequency of vocal chords
 - position of tongue, lips, mouth, ...
 - described by parameters
 - determined using spectral analysis of short audio segments
 - window of a few milliseconds
 - e.g., using Linear Predictive Coding (LPC)
 - for the next window, only store difference to previous ("frequency x increased by y")
- both, wave form and parameter encoding realized by special hardware



Music: MIDI

- „Music Instrument Digital Interface“
 - used by music industry since 1983
 - defines an interface among (electronic) musical instruments (or computers)
- Representation based on instruments
 - type of instrument (e.g., grand piano), start/end of note, pitch, volume, etc.
 - 10 octaves, i.e., 128 notes
- requires ~200KB of MIDI-data for 10 minutes of music – a lot less than sampling
 - Keyboard → Computer: input,
Computer → Synthesizer: output
 - Sequencer: device for storing/editing MIDI data – often a PC with sequencer software
- MIDI-Standard („General MIDI“):
 - 16 channels with one synthesizer instrument each
 - 128 instruments (e.g., 0 = "Acoustic Grand Piano")
 - 3-16 polyphonic notes per channel



Audio Media Object

- Raw data
 - sequence of energy levels (amplitude)
or frequency components (Fourier analysis of a time window)
 - always in compressed format due to data volume
- Registration data:
 - resolution (bit depth):
number of different energy levels
often 256 (8 bits)
 - recording frequency (sampling rate)
 - number of channels (1 for mono, 2 for stereo, ...)
- Description data:
 - for speech: transcription as text
 - for music: transcription into musical notation or MIDI
 - structural information: pauses/silence



Audio Media Object (2)

- Operations:
 - Input - from a file in a specific format, or from a device (in real time!)
 - Output - to a file or a device (real time!)
 - Modification
 - cut/edit, similar to a recording studio;
audio position based on time, sequence number of sampled value
 - adjust volume (difficult: non-linear)
 - Analysis, aggregation
 - statistics for sample value distributions
 - finding pauses/silence (based on amplitude threshold)
 - speech recognition
 - Comparison (search)
 - pattern matching
 - equality is too restrictive, similarity measures?!
 - based on description data (e.g., text)
- Subtypes
 - spoken language (most important), music, nature sounds, machine sounds (vehicle motor), ...



Indexing and Retrieval of Audio

- easiest method: using title, file name ...
 - most popular
 - but names may be incomplete, subjective – hard to find
 - does not support searching for audio that "sounds like" another audio
- Content-based audio retrieval
 - query by example, query by humming
 - comparison of (sub-)sequence of sample values
 - not promising, does not account for differences in sample rate, resolution
 - extraction and comparison of features
 - average amplitude
 - frequency distribution



General Approach to Audio-Retrieval

- Classification
 - most common types: speech, music, sound, ...
 - classes usually are of different importance for the application
 - class information itself may be useful for application
- Specialized treatment of each class
 - each class requires different processing and indexing techniques
 - speech is the most important class, and a number of successful speech recognition techniques/systems exist today
 - e.g., speech: speech recognition and indexing of resulting text
- Queries
 - need to be classified, processed, indexed
- Retrieval
 - based on similarity of query features and stored audio document features
 - search space can be reduced to a specific class



Audio Properties and Features

- Basis for classification and retrieval
- Two forms of representation
 - **time-domain** (amplitude over time)
 - **frequency-domain** (signal strength over frequency)with different properties/features
- Additional features
 - subjective, e.g. timbre



Features in the Time Domain

- Amplitude
 - represents pressure level compared to normal pressure of a medium
 - silence = amplitude is zero
- Average energy
 - characterizes the loudness of the audio signal

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

with E as average energy, N as total number of sample values, x(n) as n-th sample

Features in the Time Domain (2)

- Zero-crossing rate
 - number of signal sign changes, in some sense the average dominating frequency of the signal

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

- with $\operatorname{sgn} x(n)$ as the sign of $x(n)$; 1 if $x(n)$ positive, -1 otherwise
- Silence ratio
 - percentage of sample values that are part of a period (!) of silence
 - two threshold values:
 - amplitude threshold, below which a signal is regarded as silent
 - number of subsequent silent sample values defining a period (or interval) of silence

Features in the Frequency Domain

- Every stable signal (i.e., periodic, without frequency changes) is the sum of sinusoidal signals of different frequencies
- Fourier-transformation of the signal
 - decomposes signal into frequency components with factors (coefficients)
 - presentation: factors over frequencies (energy per frequency in decibel dB)
 - also called **spectrum** of the signal
- Bandwidth
 - interval of the frequencies appearing in the signal
 - difference of highest and lowest frequency in the spectrum
 - only frequencies with energy > 3dB are considered
 - larger for music than for speech

Features in the Frequency Domain (2)

- Energy distribution
 - is directly apparent from the spectrum
 - frequencies with high energy level: useful for classification
 - e.g., music has more frequencies with high energy levels than speech
 - computed based on **bands of frequencies** with high or low levels of energy
 - energy per band: sum of energies of all frequencies in the band
 - e.g., frequencies above/below 7kHz to classify speech (which rarely shows frequencies above the threshold)
 - centroid (or median) frequency: the mean of the spectrum
 - lower for speech than for music
 - also called *brightness*

Features in the Frequency Domain (3)

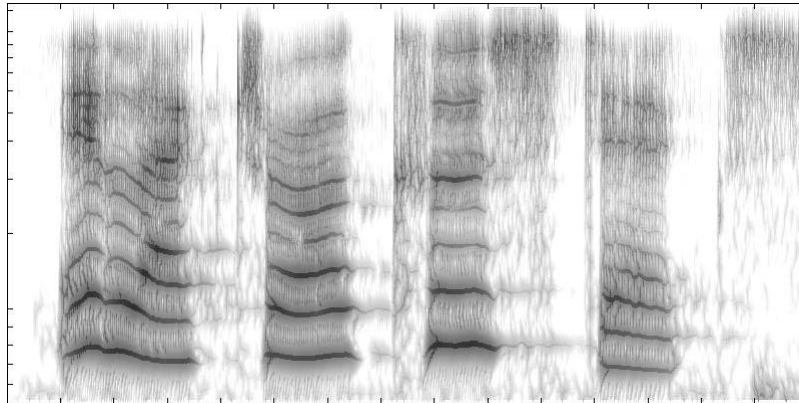
- Harmony
 - spectral components are often multiples of the lowest and loudest frequency ("fundamental frequency")
 - music is usually more harmonic than sounds
 - determining whether a recording is harmonic:
is the dominant component a multiple of the fundamental frequency?
 - example: flute plays note G₄;
 - peaks are at frequencies 400 Hz, 800 Hz, 1200 Hz, 1600 Hz
 - f , $2f$, $3f$, $4f$ etc. are harmonics of the note
- Pitch
 - defined for periodic (stable) signals (instruments, voice, ...)
 - not defined for percussion, noise, etc.
 - is subjective to human perception
 - usually close to, but not necessarily identical to the fundamental frequency)

Spectrogram

- Simple presentation has its limitations
 - time domain doesn't show frequency-related information
 - frequency domain doesn't show when the frequencies occur
- Combined presentation
 - raster image, matrix
 - x-axis: time
 - y-axis: frequency
 - blackness/intensity or color of pixel: energy of frequency at the specific time
- Analysis
 - regularity of occurrence of frequencies
 - music is more regular than speech

Spectrogram (2)

- Example
 - femal speaker, (english "Electroacoustics"), signal duration 1,5 s
 - source: <http://www.mmk.ei.tum.de/~rue/mum/eurospeech99/demo/>



© Prof. Dr.-Ing. Stefan Deßloch

19

Digital Libraries and Content Management

Classification

- Based on features
 - here only for music and speech
 - could be further differentiated:
 - types of music, male or femal speech
- Speech
 - bandwidth relatively small, 100 – 7000 Hz
 - Centroid is lower than for music
 - frequent pauses (between words, sentences) – high silence ratio
 - characteristic structure: sequence of syllables, consisting of short periods of friction (consonants) followed by longer periods of vowels – frictions show high zero crossing rate
- Music
 - high bandwidth, 16 – 20.000 Hz
 - centroid is higher
 - low silence ratio
 - except: solo instrument, a-capella singing
 - zero crossing rate does not show strong variations
 - regular beat



© Prof. Dr.-Ing. Stefan Deßloch

20

Digital Libraries and Content Management

Classification System

- Step by step, feature by feature
 - e.g., first consider centroid – if high: music
 - then silence ratio – if low: music
 - then zero-crossing variability
 - low: solo music
 - otherwise: speech
- Order is important
 - algorithmic complexity, effectiveness of classification
- A single feature is already useful:
 - only zero crossing rate: up to 90% correctness
 - only silence ration: up to 80% correctness
- Feature vector
 - combines values of a set of features
 - training: compute average vector (reference vector) of each class
 - for new audio, compute feature vector and determine distance to all reference vectors (using, e.g., euclidean distance)

Speech Recognition

- Performed after classification
- Techniques
 - Time Warping (speed of speech)
 - Hidden Markov Models
 - neural networks
- Performance

<i>domain</i>	<i>type</i>	<i>vocabulary</i>	<i>error rate in %</i>
digits	read	10	< 0,3
flight reservation system	spontaneous	2500	2
Wall Street Journal	read	64000	7
radio news	read / spontaneous	64000	30
phone call	spontaneous	10000	50

Music Indexing

- Structured music (MIDI)
 - no extraction of features required
 - exact match is a valid search option
 - but maybe the instrument is different for some tracks
 - similarity is hard to define
 - one option: only consider change of pitch
 - Up, Down, Repeat – U, D, R
 - Parsons, D., *The Directory of Tunes and Musical Themes*, Spencer Brown, 1975
 - Melodyhound: <http://name-this-tune.com/> (Uni Karlsruhe)retrieval reduced to character string comparison
- Recorded music (sample-based)
 - Query: singing, humming
 - Set of features to consider
 - e.g., volume, pitch, brightness, bandwidth, harmony
 - vector and distance computation
 - Pitch
 - extract/guess for each note ("pitch tracking")
 - representation as a series of pitches, or pitch changes (see above)
 - similarity: sequences may differ in k pitches

Summary

- Audio Digitization
 - PCM
- Audio Encoding
 - Waveform Encoding
 - Parameter Encoding
- Music: MIDI
- Audio Media Object
- Indexing and Retrieval of Audio
 - general approach: classification-based
 - features in the time domain: energy, zero-crossing rate, silence ratio
 - features in the frequency domain: bandwidth, energy distribution, brightness, fundamental frequency, harmony, pitch
 - spectrogram combines time and frequency domains
- Classification
- Music Indexing: still not mature enough