VK Multimedia Information Systems

Mathias Lux, mlux@itec.uni-klu.ac.at

Dienstags, 16.oo Uhr c.t., E.1.42

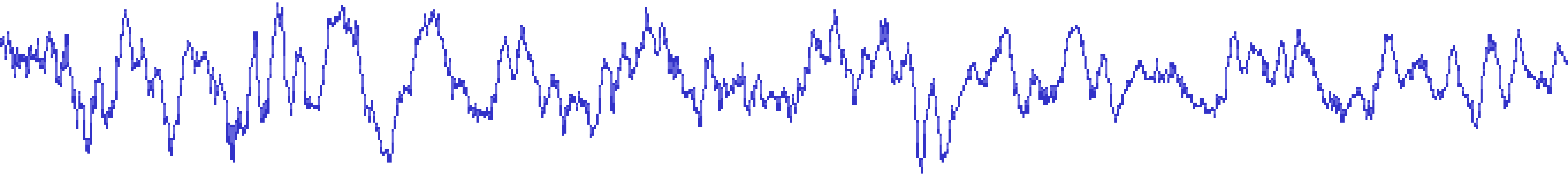
Audio & Music Retrieval

- What is Digital Audio?
- Features & Descriptors
- Speech
  - Speech Recognition
  - Speaker Detection
- Event Detection
- Music Retrieval
  - Motivation & Problems
  - Algorithms & Methods

What is Digital Audio

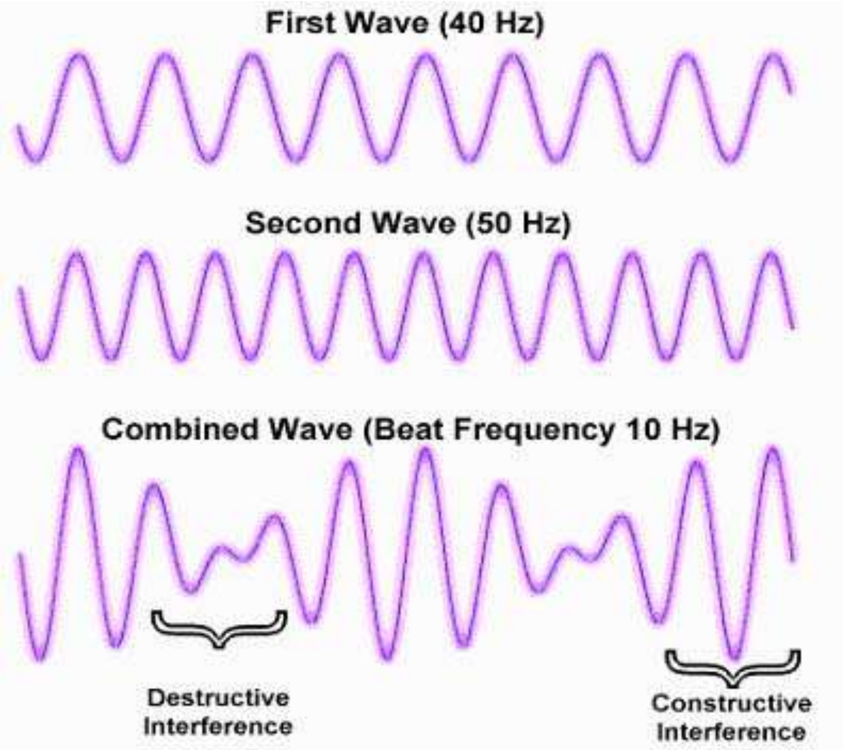
- Analogue signal goes digital
- Digitization: PCM
- Formats:
  - Compression
  - Containers

What is sound?



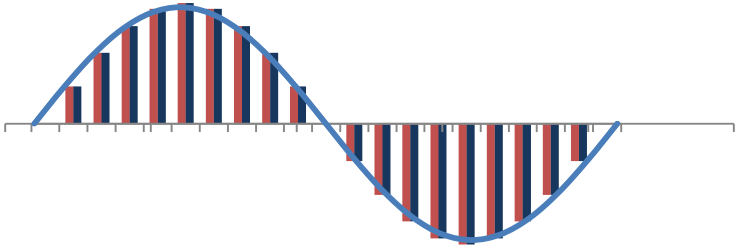
What is sound?

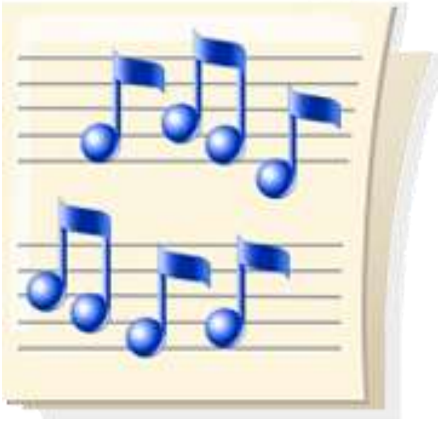
- Multiple sounds at the same time?


First Wave (40 Hz)
Second Wave (50 Hz)
Combined Wave (Beat Frequency 10 Hz)
Destructive Interference
Constructive Interference

What is digital sound?

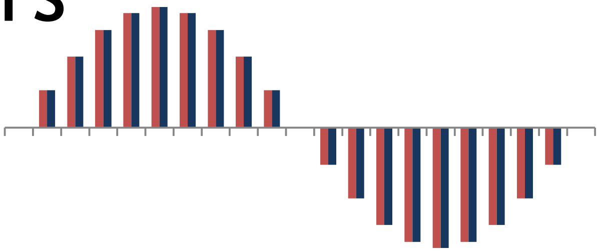
- A digitization of the wave.
  - – Either a recipe for reconstruction
  - – Or a discrete approximation





Sampled sound

- Wave gets sampled x times a second
  - E.g. 48.000 times -> 48 kHz sampling rate
- Obtained values are stored
  - E.g. 256, 240, 13, -7, -12, -44, ...
  - Quantization to e.g. 2^8 levels -> 8 Bit
- Possibly from different sensors
  - Stereo -> 2 channels

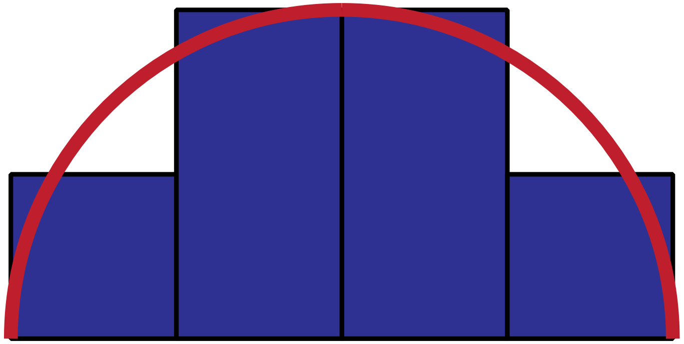


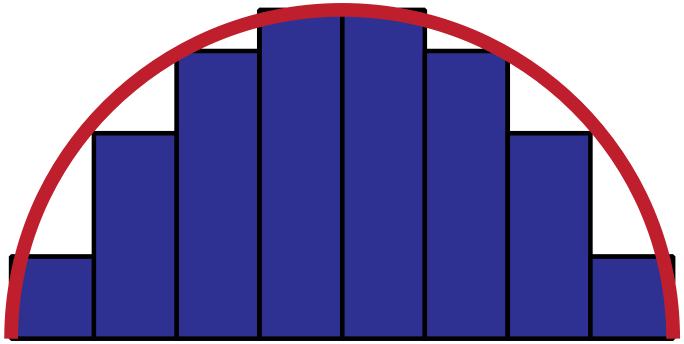
Sampled sound

- Example: 8 kHz, 16 bit Stereo
  - Sound wave is sampled 8.000 times a second
  - Samples are stored in 16 bit numbers
- That's *Pulse Code Modulation* (PCM)
  - Often used in WAV files …
  - Also as input from microphone or line in

Sampling Rates

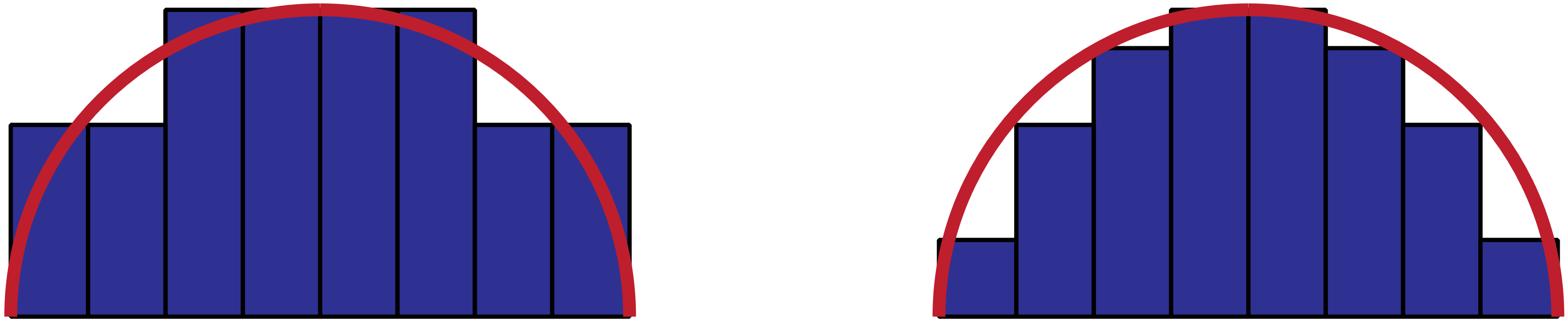
- With sampling rate x we can approximate frequencies up to x/2
- Assume frequency 1
  - – sampling rate of 1 -> “0”
  - – sampling rate of 2 -> “1,-1”





Quantization

- Reduces the possible values of the samples to a certain value
  - 8 Bit -> 256 levels, etc.



What do we want to capture?

- Humans can hear
  - From around 16 – 21 Hz
  - To around 16 kHz – 19kHz
  - 16 bit is enough (CD), 32 bit even better

Sound Formats

- Waveform Audio Format
  - Container for several compression formats
  - Includes PCM, MP3, GSM, μ-Law
- Musical Instrument Digital Interface
  - Control codes for instruments
  - Instruments can be “emulated”
- Compressed Audio Formats
  - MP3, OGG, AAC, …

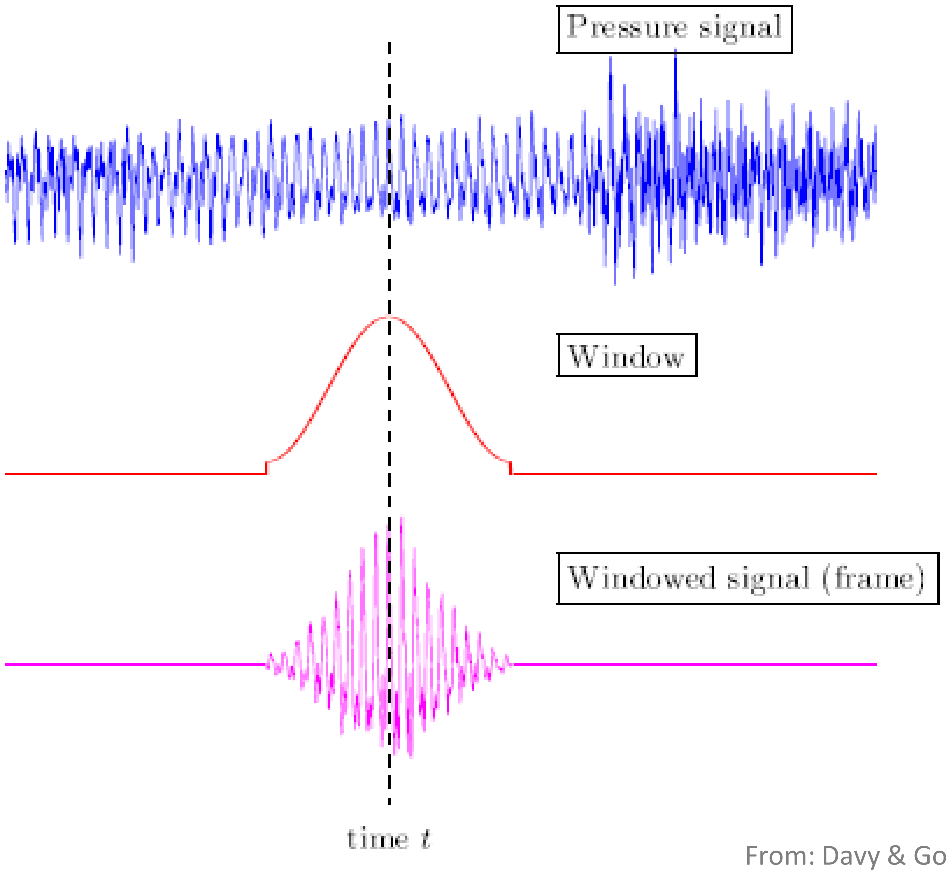
Audio & Music Retrieval

- What is Digital Audio?
- **Features & Descriptors**
- Speech
  - Speech Recognition
  - Speaker Detection
- Event Detection
- Music Retrieval
  - Motivation & Problems
  - Algorithms & Methods

Audio Low-Level Features

- Lots of different applications
  - Speech Recognition
    - One of the first applications
  - Music Information Retrieval
  - Environmental Sound Recognition
- Different specific Features / Descriptors
- Standardization Efforts
  - MPEG-7, *low-level audio descriptors*

Audio Frames



From: Davy & Goodsill, " Audio Information Retrieval: A Bibliographical Study", TR, 2002

Audio Low-Level Features

- Features describe audio frames
- Frame definition critical to outcome
  - Shape (rectangular, Hamming, etc.)
  - Size (e.g. 150 ms)
- Features capture aspects
  - Energy (loudness)
  - Frequencies
  - Change over time (attack time, etc.)

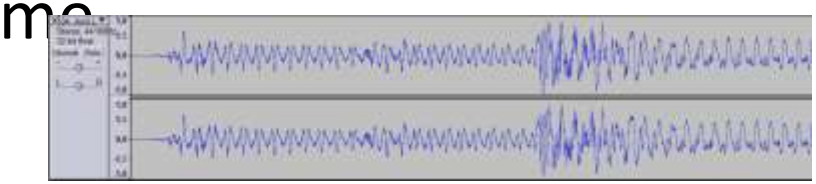
Audio Low-Level Features

- MPEG-7 collects assortment of usable features
  - Basic
  - Basic Spectral
  - Spectral Basis
  - Signal Parameters
  - Timbral Temporal
  - Timbral Spectral

Basic Features

Short description of audio waveform

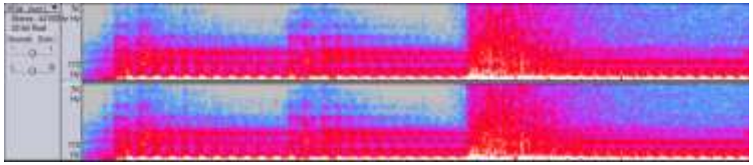
- AudioWaveform
  - Mainly for display
  - Minimum and maximum of the envelope
- AudioPower
  - Average square of the waveform samples
  - Power of the signal over time



Basic Spectral

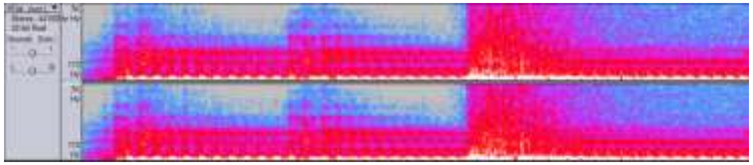
Basic properties of the spectrum of a signal

- AudioSpectrumEnvelope
  - short-term power spectrum of a signal
  - logarithmic frequency scale
  - imitating the human ear
- AudioSpectrumCentroid
  - Center of gravity in above descriptor
  - Indicates whether high/low freq. dominate



Basic Spectral

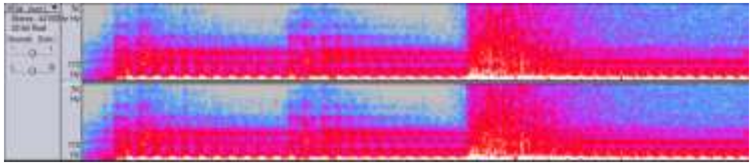
- AudioSpectrumSpread
  - – deviation of the power spectrum from centroid
  - – separation of tonal from noise-like sounds
- AudioSpectrumFlatness
  - – deviation of the spectrum from a flat shape
  - – designed to perform *fingerprinting*



Spectral Basis

… general-purpose sound recognition

- AudioSpectrumBasis
  - Transforms spectrum to a lower-dimensional representation
  - Based on power spectrum
- AudioSpectrumProjection
  - Also transformation / reduction of data
  - Based on orig. signal & above descriptor



Signal Parameters

- AudioFundamentalFrequency
  - Fundamental frequency of a sound
  - Applicable to sound segmentation of speech and music
- AudioHarmonicity
  - Measure for the degree of harmonicity
  - Allows distinction between
    - sounds with a harmonic spectrum (e.g., musical tones or voiced speech [e.g., vowels]),
    - sounds with an inharmonic spectrum (e.g., metallic or bell-like sounds) and
    - sounds with a non-harmonic spectrum (e.g., noise, unvoiced speech, or dense mixtures of instruments)

Timbral Temporal

- Usually employed in music retrieval, independent of pitch and loudness
- LogAttackTime
  - logarithm of attack time of a sound
  - attack time is the time from the beginning of a sound signal to a point in time where its amplitude reaches a maximum
- TemporalCentroid
  - Time point of highest signal energy

Timbral Spectral

Descriptors rely on harmonic peak estimation

- Harmonic peaks
  - Correspond to frequencies that are a multiple of the fundamental frequency
  - Are used to describe the timbre of a signal

Timbral Spectral

- HarmonicSpectralCentroid
  - Amplitude-weighted average of harmonic peaks in spectrum
- HarmonicSpectralSpread
  - Amplitude-weighted deviation of harmonic peaks from above feature

Timbral Spectral

- HarmonicSpectralDeviation
  - Deviation of harmonic peaks from spectral envelope
- HarmonicSpectralVariation
  - Correlation of harmonic peaks in adjacent frames
- SpectralCentroid
  - Power-weighted average of frequencies in the power spectrum

Audio & Music Retrieval

- Features & Descriptors
- **Speech**
  - Speech Recognition
  - Speaker Detection
- Event Detection
- Music Retrieval
  - Motivation & Problems
  - Algorithms & Methods

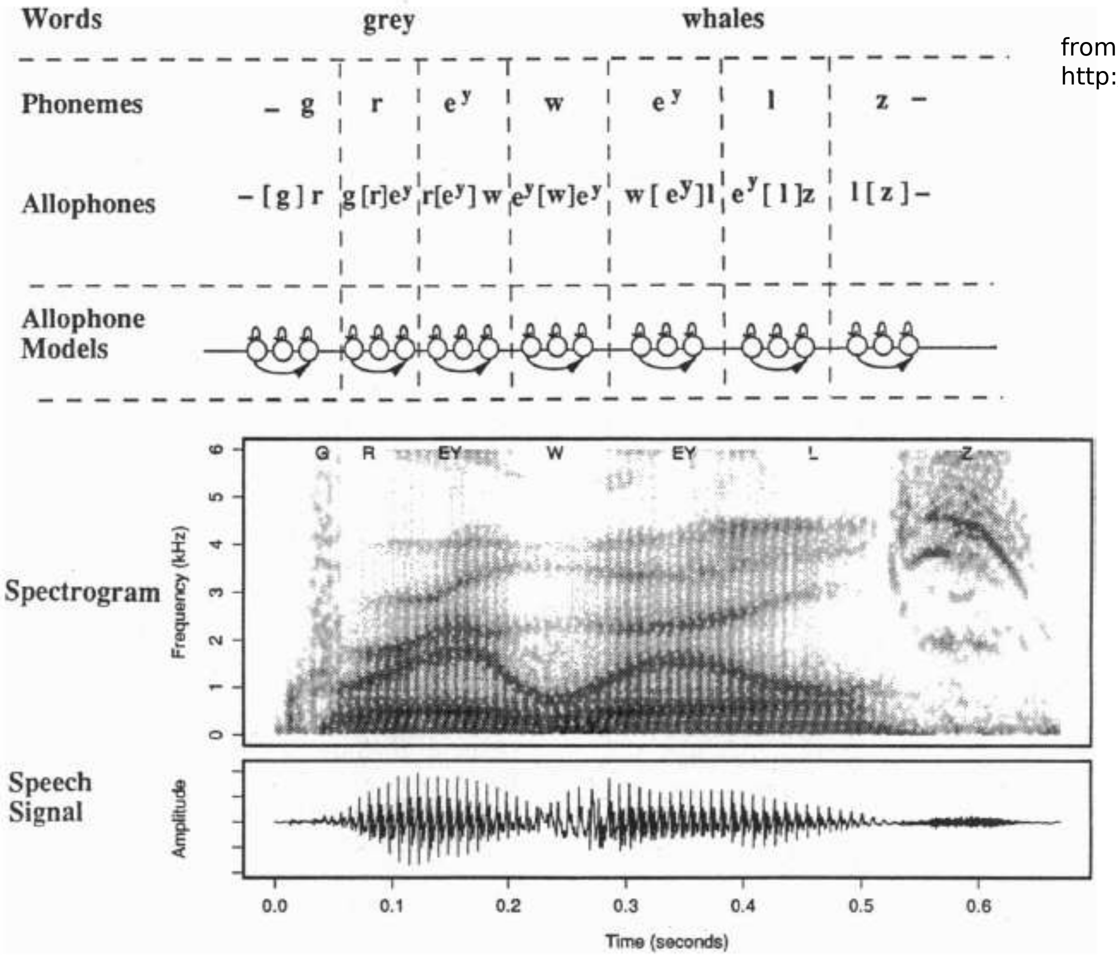
Speech Recognition: The Problem

- Conversion of **acoustic signal** into **words**
- Different possible approaches
  - Isolated-word speech recognition
  - Continuous speech recognition
- Dependence on speaker
  - Training samples or independent

Speech Recognition: The Problem

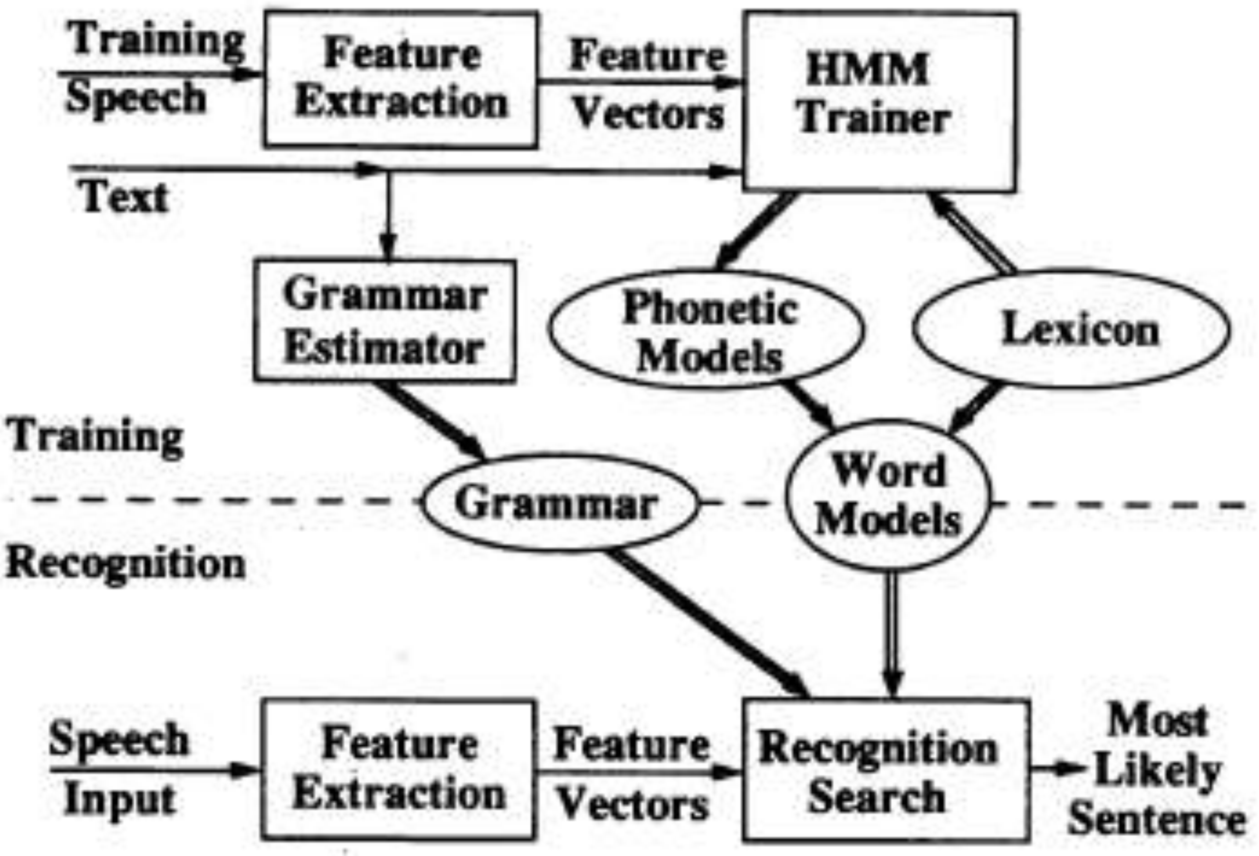
- Spontaneous vs. speech read from script
  - Spontaneous has disfluencies, more challenging task
- Language model is used for word sequences
  - Restriction to combination of words
- Measure for difficulty of the task: *Perplexity*
  - geometric mean of the number of words that can follow a word after the language model has been applied

Words & Phonemes



from: State of the Art in Continuous Speech Recognition: http://www.pnas.org/cgi/reprint/92/22/9956

Speech Recognition: Process



Speech Recognition: Process

- Feature Extraction
  - Recognition from signal (near real time)
  - Amount of data (matching & indexing)
- Training
  - Modeling characteristics of speakers
  - Pronounciation -> Phonetic HMM
  - Grammar -> Markov Model (bi- or tri-word)
- Recognition
  - Search among possible word sequences
  - Highest possibility as match

Speaker Recognition

Detect speaker (change) in continuous speech

- Proper features to describe speaker
  - based on group of possible speakers
- Proper classification algorithm
- Robust against natural influence
  - Noise, cold, emotions, etc.

Speaker Recognition: Applications

- Video Analysis
  - – Segmentation of interviews, etc.
- Media Analysis
  - – How long did the J. Doe speak in TV this month?
- Security
  - – Access restrictions: “Computer, shut down the warp drive!”

Audio & Music Retrieval

- Features & Descriptors
- Speech
  - Speech Recognition
  - Speaker Detection
- **Event-Detection**
- Music Retrieval
  - Motivation & Problems
  - Algorithms & Methods

Event Detection

- Several domains have simple characteristics:
  - **Sports** events follow rules, participants behave similar
  - **News** broadcasts have a simple scheme, news anchormen introduce and summarize stories
  - **Surveillance** applied in ‘dull’ scenarios to detect ‘extreme situations’ like fire, panic, etc.
  - **Ad blocks** in TV have rough and fast scene cuts and raised volume

Event Detection

- Events are ‘peaks’ in one or several dimensions
  - Appropriate dimensions have to be found
  - Possibility of event has to be calculated
- Several domain rules might apply
  - Scoring in soccer after final whistle not possible
  - Foul and applause unlikely to occur immediately after another
  - Certain temporal distance between ad blocks

Event Detection: Example Soccer

- Event of Scoring:
  - Applause following a goal
  - Raised volume in commentators voice
  - Whistle of the referee
- Event of Foul:
  - Cheers of ‘boo’
  - Whistle of the referee
- Event of Start / End of game
  - First and last whistle of the referee
  - Certain minimum amount of time in between

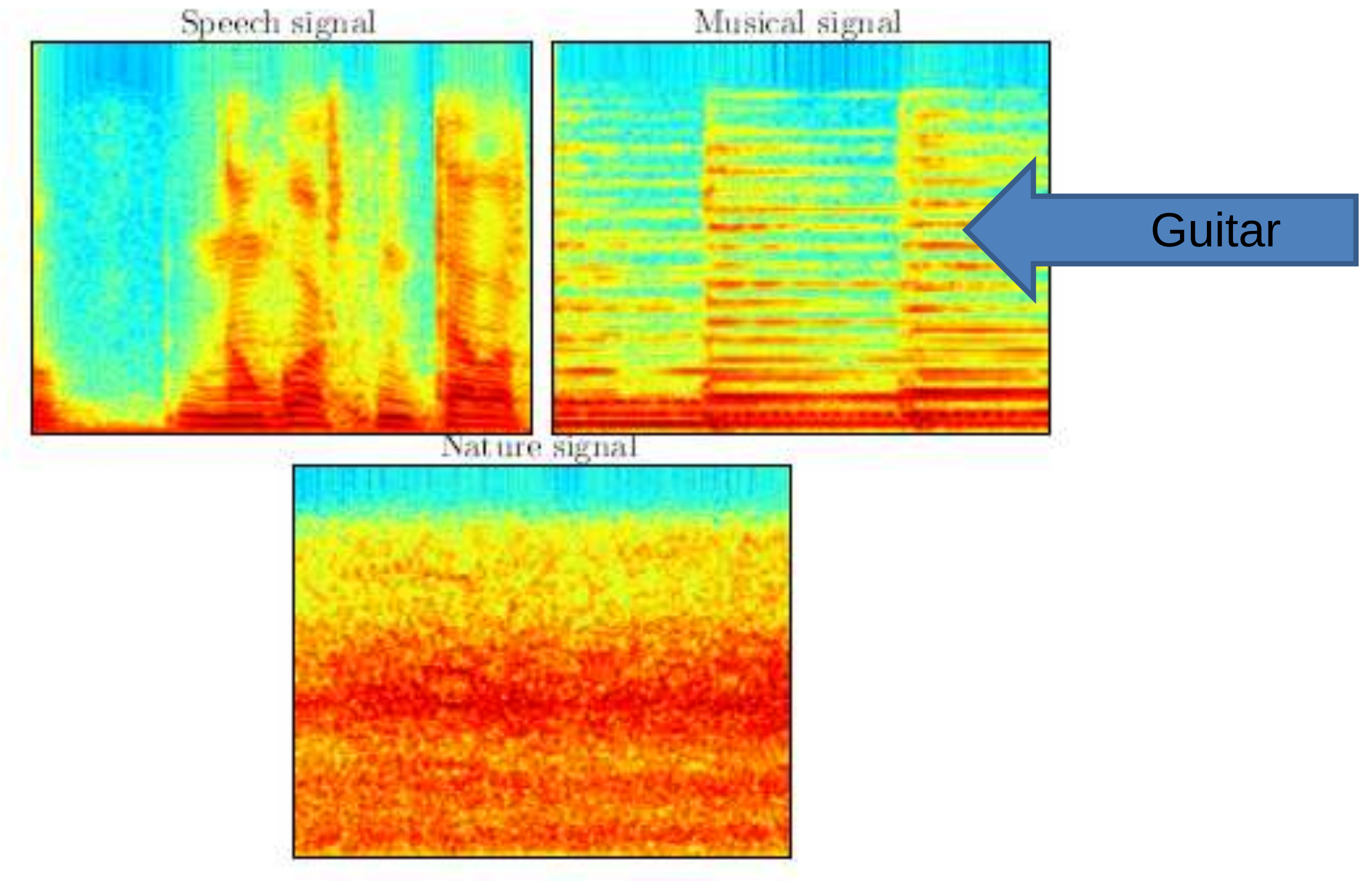
Audio & Music Retrieval

- Features & Descriptors
- Speech
  - Speech Recognition
  - Speaker Detection
- Event Detection
- **Music Retrieval**

MIR: Background

Sounds produced by musical instruments are …

- almost **periodic vibrations**
- a combination of **different frequencies**
- all multiple integers of a **fundamental frequency** (called *F0*)


Speech signal
Musical signal
Nat ure signal
Guitar

MIR: Three basic features

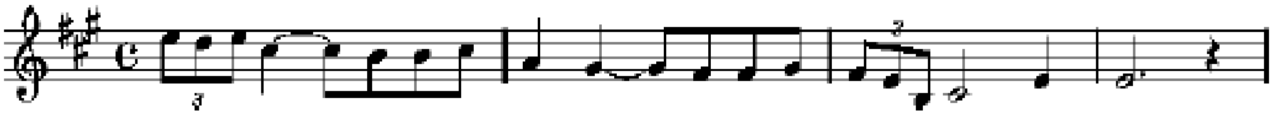
- **Pitch**
  - based on the fundamental frequency
  - low/deep to high/acute
- **Intensity**
  - the intensity of the amplitude
  - the energy of the vibration
- **Timbre**
  - sound characteristics that allow listeners to perceive as different two sounds with same pitch and same intensity

The User

- Three different intentions
  - Listen to particular performance / musical work
  - Building a collection of music
  - Verifying or identifying works
- Information need of users
  - Formalization of need often not easy
    - Unexperienced vs. professional user
  - Query-by-example easier
  - Possible task: Automatic playlist generation

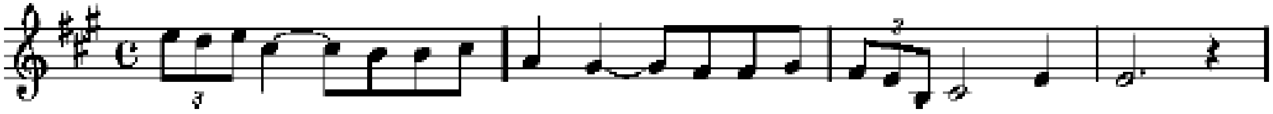
Music Processing: Melody

- Melody is key feature
  - Many genres have single relevant melody line
  - Discrimination even without rhythm
  - Eventually easy to extract (e.g. Midi)
  - Main feature for query-by-humming
- Retrieved & indexed using n-grams
  - Short sequences of same length
  - Segmentation remains issue



Music Processing: Melody

- Extraction
  - Pitch Tracking
  - F0-Estimation
- Limitations
  - Single note vibrato



Music Processing: Harmony

- Chord sequences are considered as relevant descriptors
- Extraction is challenging task
  - Transcription even for user hard

Music Processing: Timbre

- Most difficult feature to characterize
  - defined as acoustic feature that is neither pitch nor intensity
  - Mainly related to the spectrum
- Timbre parameters are left to choice of performers
  - At least in Western classical music
  - Not formalized through transcription
- Listeners are very sensitive to change in timbre

Music Processing: Orchestration

- How the particular work is orchestrated
- Described through musical instruments
  - Style of play is more a matter of timbre
- Recognition of musical instruments
  - Main way of extraction
  - Used rarely in MIR
  - Recognition rates rather high
    - 100% for easily recognizable instruments (like flute)
    - 75% for harder tasks (like chello)
    - 80% on average on large instrument databases

Music Processing: Rhythm

- Intuitively easily recognizable
  - Assumption comes from Western music
  - Africa & Eastern Europe -> highly complex task
- Pop and Rock music simple examples
  - Rhythm is based on variations
  - Four equally spaced beats
  - 1st and 3rd are stronger
- Tempo Tracking
  - relevant for dance / mix / radio

Music Information Retrieval: Examples

Pandora … www.pandora.com

- Pandora is a personalized internet radio
  - Selection of 'songs one likes'
  - Stream composed on music retrieval
- Audio Genes (fingerprinting)
  - Retrieve music with similar content
- Recommenders and Classifications
  - Based on several characteristics like heavy guitars, impressive voice, etc.

Music Information Retrieval: Musipedia.org

- Portal for searching music
- Several different search options
  - Keyboard search
  - Contour search
  - Query by humming
  - Rhythm search
- Example: Contour search
  - U ... Up, D ... Down, R ... Repeat
  - DDUUUDRDR - Austrian National Anthem
  - UUDUDDDUUDDDDUDU – Haydn, Emporer’s Hymn

Thank you ...

... for your attention