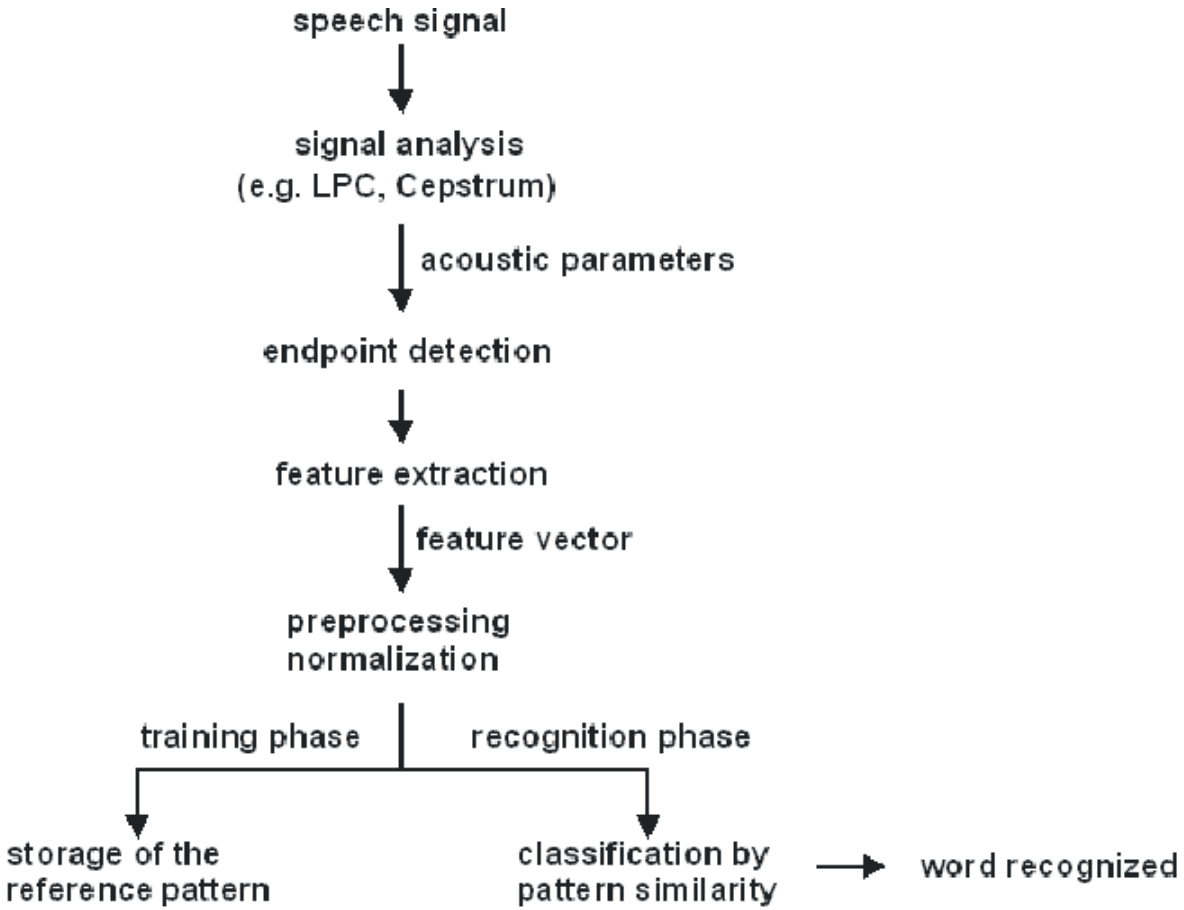
8.5 Deriving Audio Semantics

8.5.1 Speech Recognition

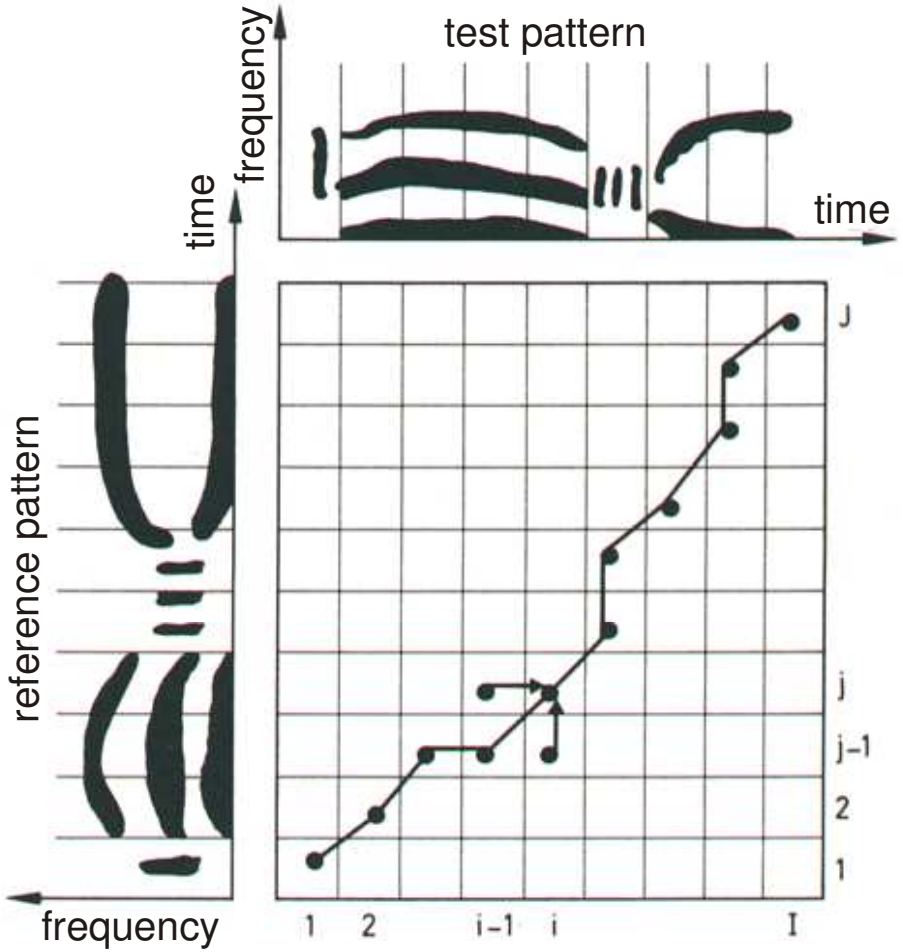


Speech Recognition Technology (1)

Word Matching (Dynamic Time Warping)

Problem: For the same word, timing is different in different speech situations.

We can use dynamic programming to map two versions of a word onto each other. Dots in the diagram identify identical patterns.



Speech Recognition Technology (2)

Hidden Markov Models (HMMs)

The process of speech generation is modeled by a stochastic finite automaton. The automaton computes the probability of having itself generated a given speech signal. In a **learning phase**, a separate automaton is trained for each word of the language. In the **recognition phase** the automata then compute their probabilities.

Reference

J.R. Deller, J.G. Proakis, J.G.H. Hansen: Discrete-Time Processing of Speech Signals, Prentice Hall 1987

Speaker Recognition

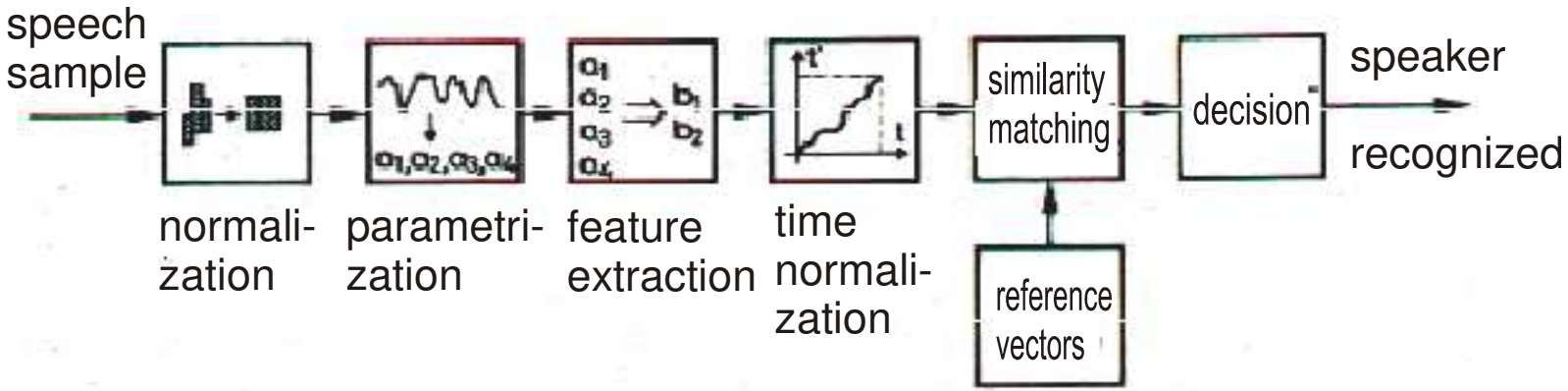
Speaker recognition is much easier than speech recognition. We try to match the parameters of a given speech sample with parameters contained in a database.

Application examples

- Identification of criminals
- Access control for buildings

Method

- Pattern recognition



Individual Speech Parameters

How speech sounds depends on different dimensions of the vocal tract and the vocal cords as well as on speech behavior learned over many years. The latter can be faked easily so that the former is better suited for speaker recognition.

Method

- Computation of long-term averages and standard deviations of speech features
- Computation of histograms for the features

8.5.2 Silence Detection

Goal

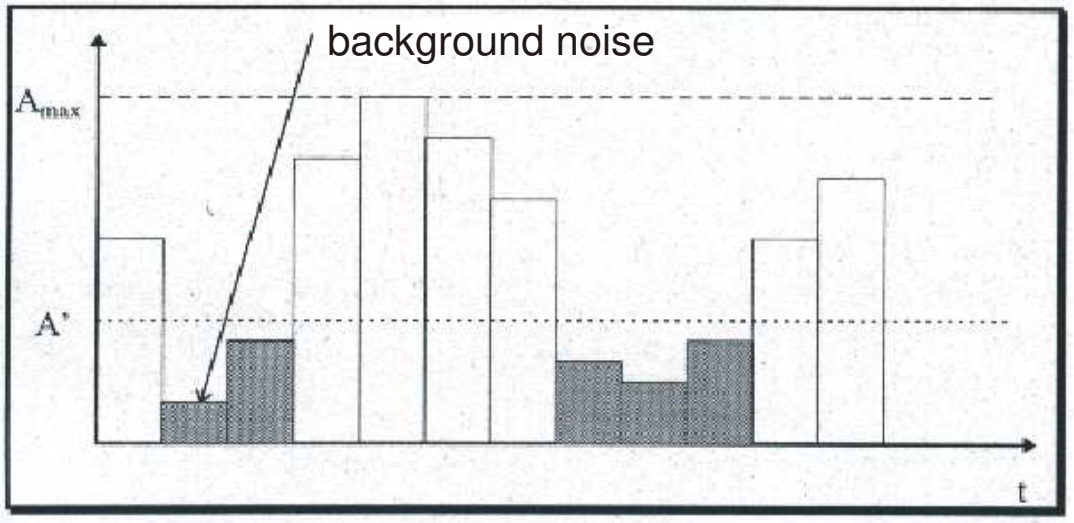
Detection of (relative) silence in the audio track.

In natural sound situations there are moments which the human listener identifies as silence. They a characterized by the lack of a dominant foreground sound (such as speech). A low-level background sound might be present.

Silence detection techniques

1. Measurement of loudness

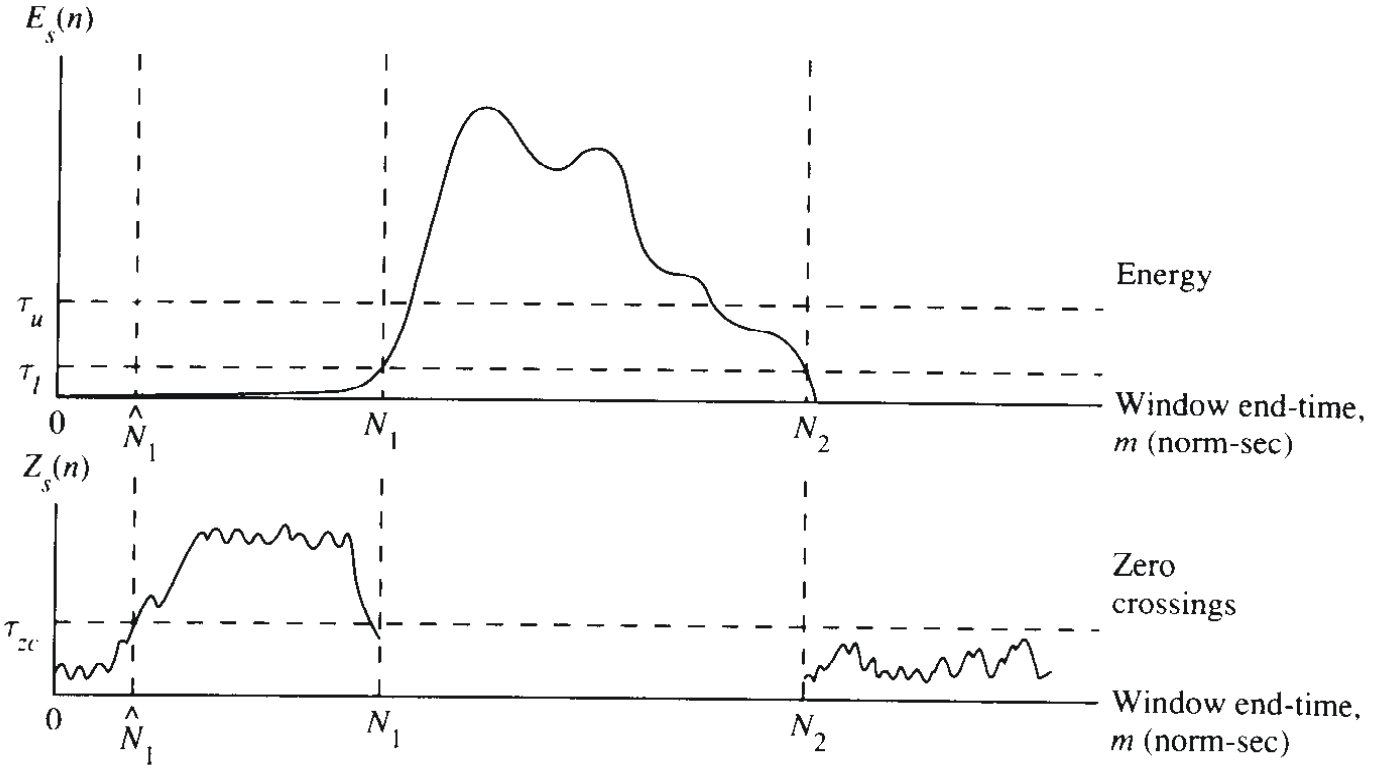
Phases with small loudness values are marked as silence.



Silence Detection (2)

2. Measurement of signal energy

Often used in combination with zero crossings to detect word boundaries in speech.

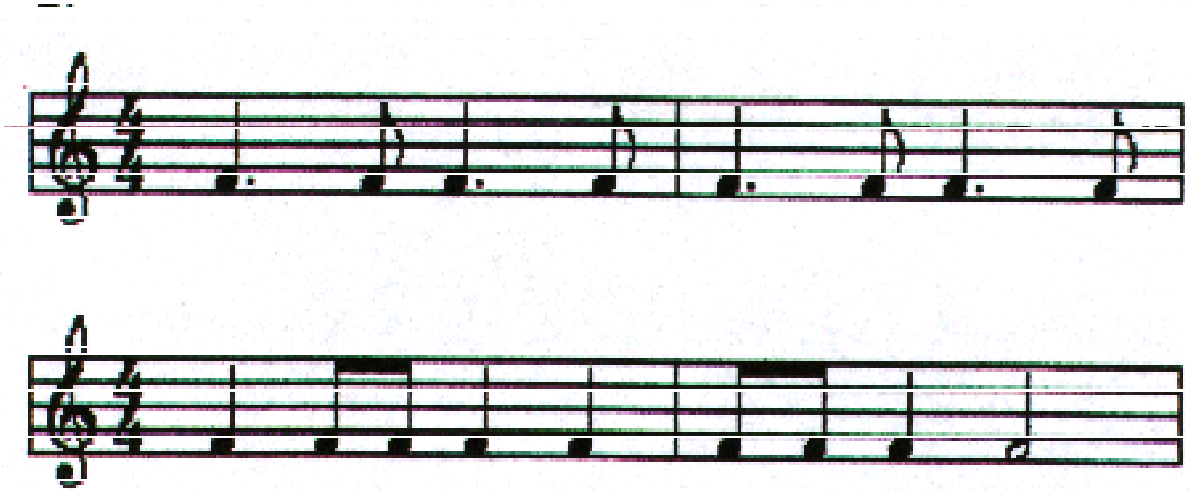


8.5.3 Temporal Structure

Time and rhythm in music

Time = regular pattern of emphasized and non-emphasized sounds, e.g.: ¾ time for a waltz.

Rhythm = characteristic sequence of beats that can span several bars

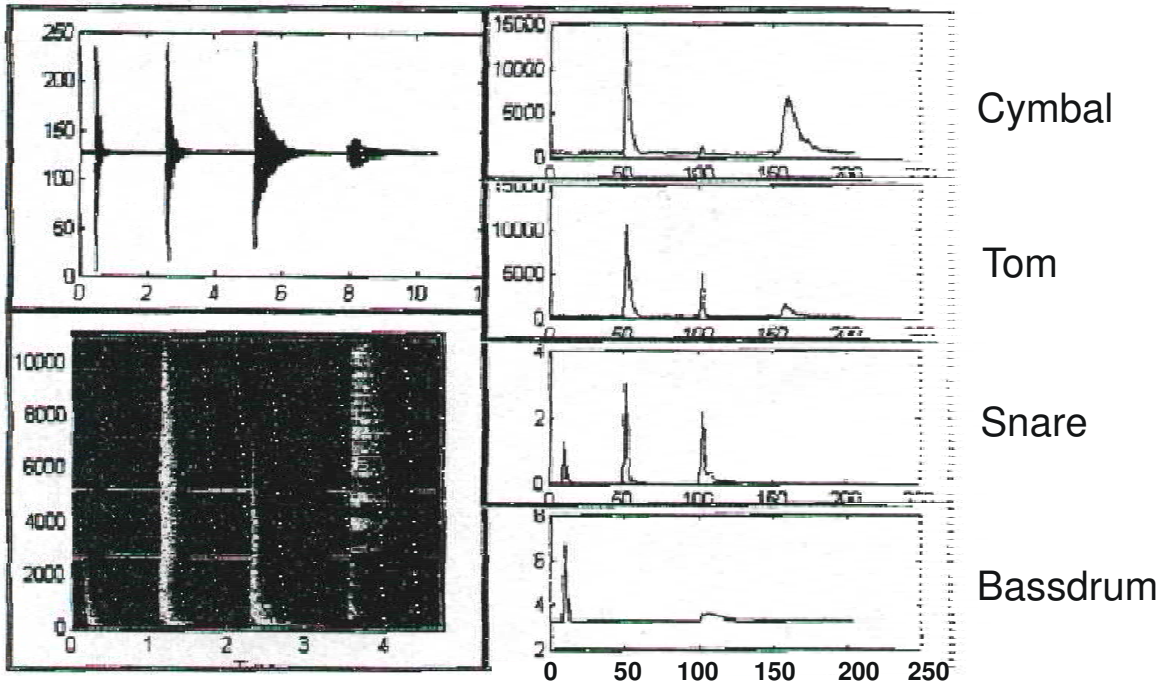


Time and Rhythm Detection

In the time domain: statistics on amplitudes

- Detect peak amplitudes (absolute or relative) and use a threshold
- Empirical result: does not work because peaks cannot be identified reliably in music

In the frequency domain:
try to identify
the drums



Recognizing drums with the FFT Tool

Rhythm Detection: Results

In an empirical evaluation done at U. Mannheim we were able to detect the rhythm in 15 out of 20 pieces of music.

Identification of Instruments

Important parameters

- Spectral distribution of energy
  - Frequency spectrum of the instrument
  - Specific harmonics
- Temporal structure of the frequency components
  - at the beginning (onset)
  - in the stable phase
  - at the end of a tone.
- Transitions between tones

For example, the frequency spectrum of brass instruments remains fairly stable in the middle phase of a tone whereas it varies considerably for string instruments (vibrato). For a trumpet, the onset is short whereas for a clarinet or saxophone, the onset is longer.

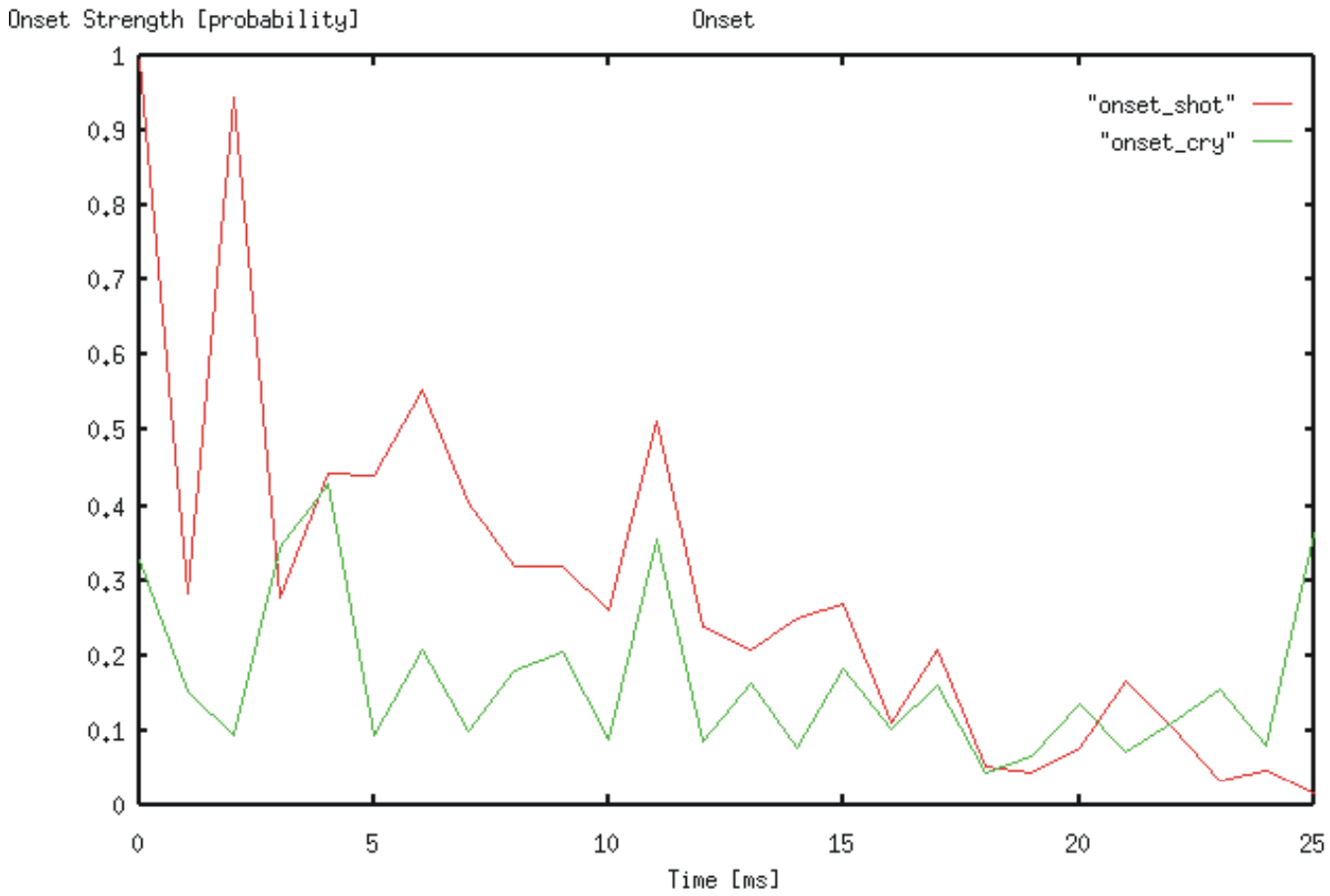
Onset of a Sound

Example

A gun shot

vs.

a scream



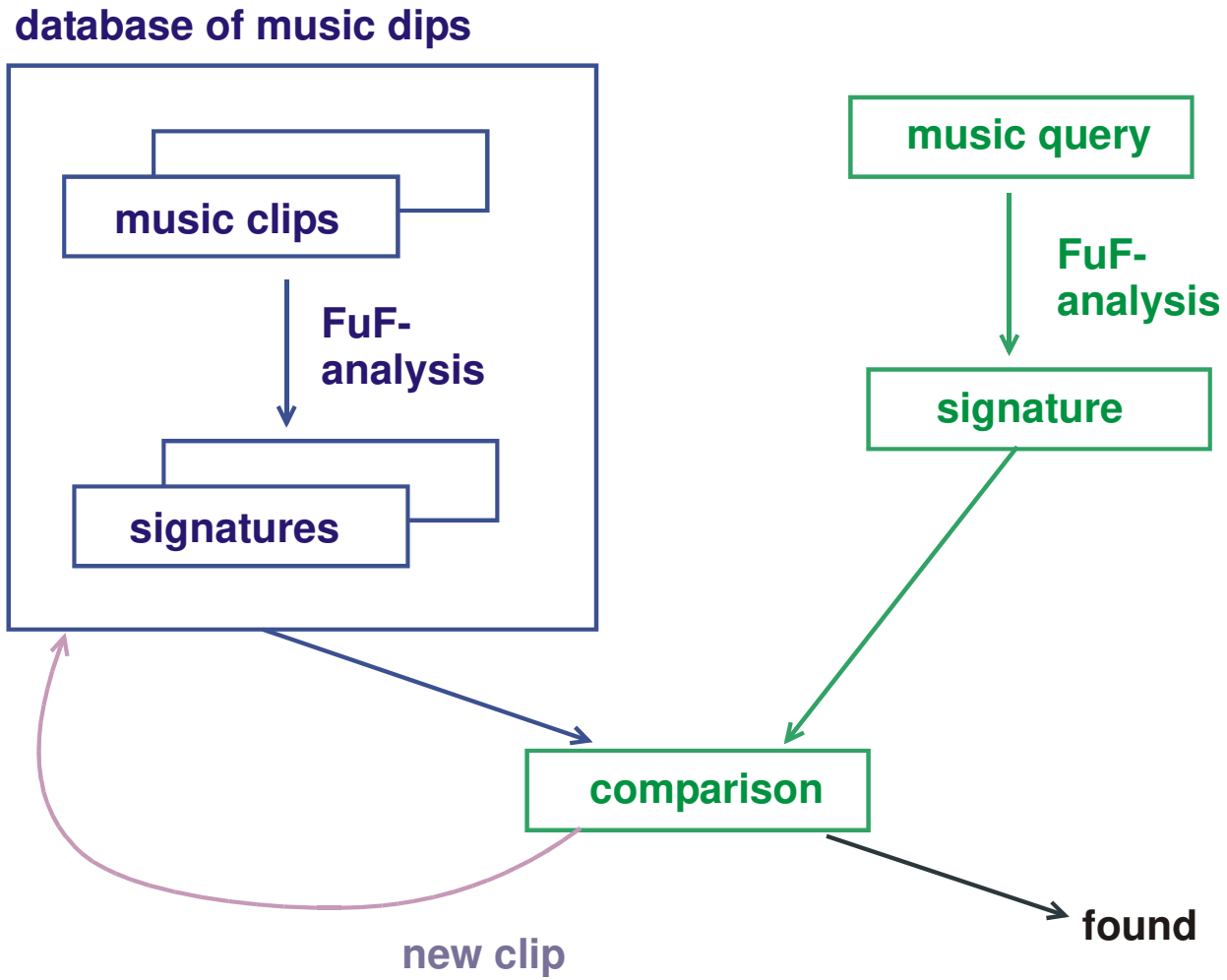
8.5.4 Application Examples

- Distinction of silence, speech, music and noise on the audio track of a video
- Transcription of speech to ASCII for subsequent automatic indexing of newscast (done in the Informedia project of Carnegie Mellon University)
- Full understanding of audio semantics in specific cases: tennis, baseball, shots, explosions, animal sounds (the barking of a dog)
- Detection of important scenes in sports events based on the audio level of the spectators
- “query by example“ in large music databases
- Detection of copyright violations for music on the Internet (finding stolen audio clips)

Search in Music Databases (1)

Music clips can be stored as waveforms or in MIDI format.

For **waveforms** we can use *parameter signatures* to compare clips.



Search in Music Databases (2)

For music clips stored in **MIDI format** we can use melody matching techniques. They allow the abstraction from details in their definition of similarity.

Method

- Describe a melody as a sequence of UP, DOWN or SAME tones. Ignore the duration of the tones.
- Use string matching techniques to compare the sequences.
- Example: *SUUSDDDDSUUSDS

Reference

A. Uitdenborgerd, J. Zobel: Melody Matching Techniques for Large Music Databases, Proc. ACM Multimedia 1999, Orlando, Florida, pp. 57-66