

Human sound perception — what can we learn from it when developing audio analysis algorithms?

Tuomas Virtanen
tuomas.virtanen@tut.fi

Tampere University of Technology, Department of Signal Processing

7th September 2012

Why perceptually motivated methods?

- ▶ *“I already tried auditory model X for feature extraction and it did not improve the results — why bother?”*
- ▶ In most computational audio analysis tasks, we are still far from the capability of human perception

→ We should see what we can learn from the human audio perception

Why perceptually motivated methods?

- ▶ Most of “perceptually motivated” methods take ideas from the early processing stages of the human auditory system
- ▶ Ideas successfully used in specific tasks (source separation, f_0 estimation, etc.)
- ▶ Perception involves also high-level processing, which is not currently utilized properly in computational methods

The purpose of this talk

- ▶ Add some perspectives about high-level processing
- ▶ To discuss the properties of perception more broadly, application independently

Outline of the talk

Motivation and purpose of the talk

High-level properties of human audio perception

Models of human perception

Peripheral processing

Bottom-up processing

Top-down processing

High-level properties of human audio perception

The goal of human audio perception: to extract information from our surroundings by means of sound = auditory scene analysis

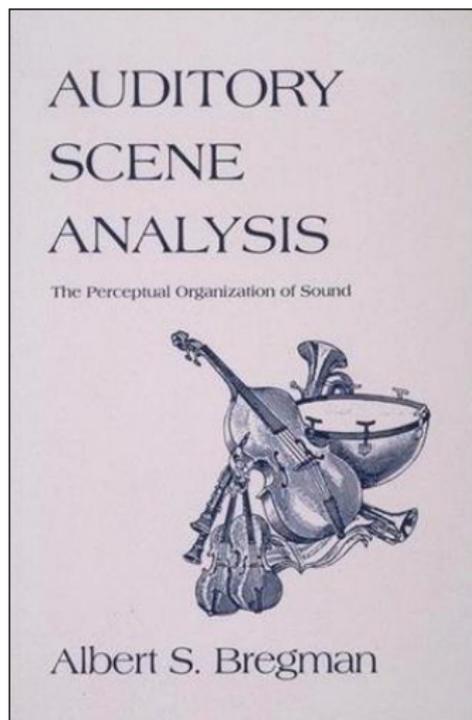
- ▶ Audio perception: ability to construct a mental, symbolic representation of a sound, based on an acoustic input
- ▶ Scalability: ability to do the above reliably for a large number of different types of sounds
- ▶ Generalization: ability to do the above even for sounds that we have not heard before
- ▶ Robustness: ability to do the above in diverse conditions
- ▶ Adaptivity: ability to learn efficiently models for new types of sounds, ability to adapt to changes
- ▶ Source separation: ability to perceive a sound in a mixture of other sounds
- ▶ Spatial hearing: ability to perceive the location of a sound source, and to estimate the properties of the environment

Sources of information

- ▶ Psychoacoustics
- ▶ Auditory neurophysiology
- ▶ Audiology
- ▶ Cognitive sciences — auditory cognition
- ▶ Neuroimaging
- ▶ Neuroinformatics

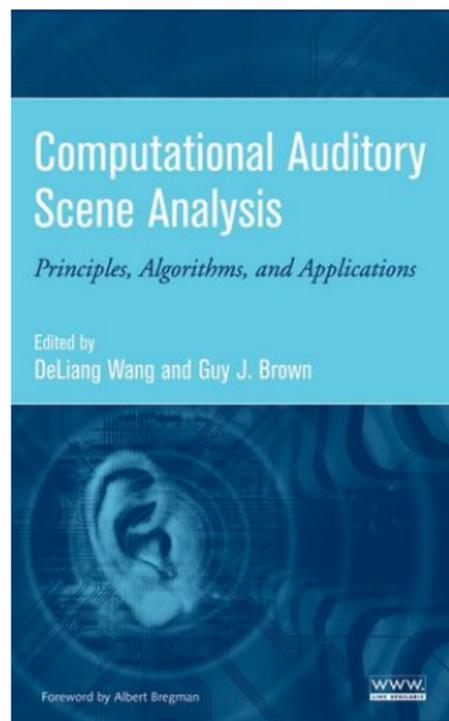
The auditory scene analysis book

- ▶ Book by Albert S. Bregman, published in 1990.
- ▶ Describes the principles that the human perception uses to organize sounds
- ▶ Initiated plenty of computational auditory scene research in late 1990s and early 2000s

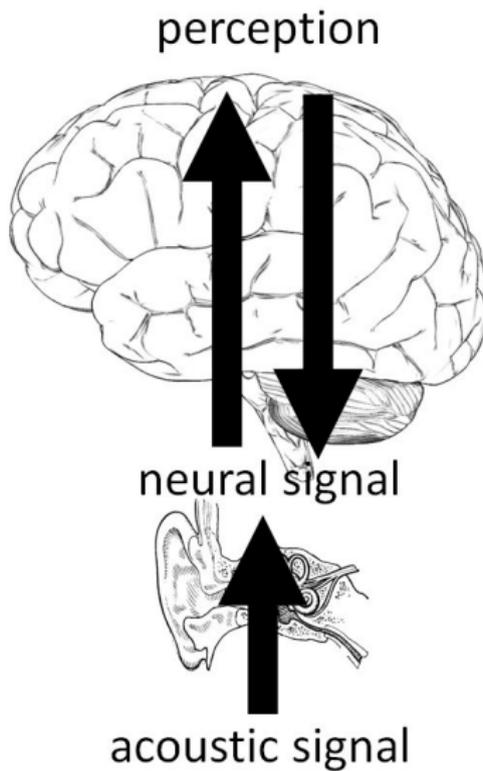


The CASA book

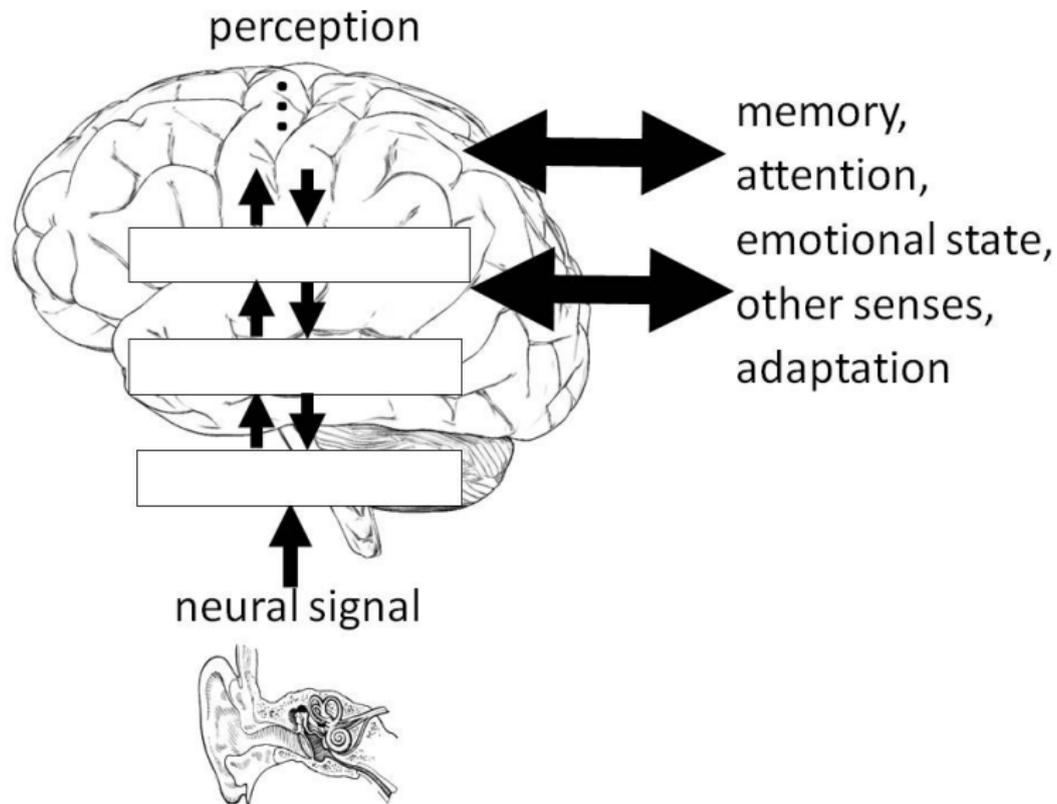
- ▶ CASA = computational auditory scene analysis
- ▶ Book edited by DeLiang Wang and Guy J. Brown, published in 2006
- ▶ Collected the main results related to CASA



A simplified figure of signal processing in the auditory system



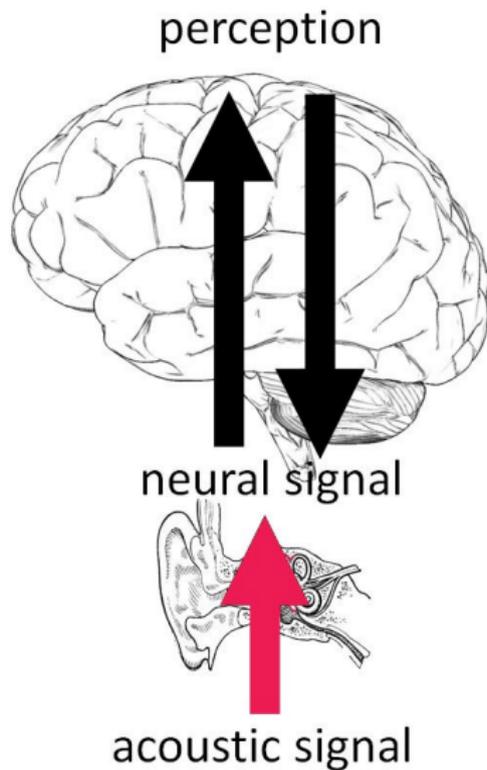
More extensive view of signal processing in the auditory system



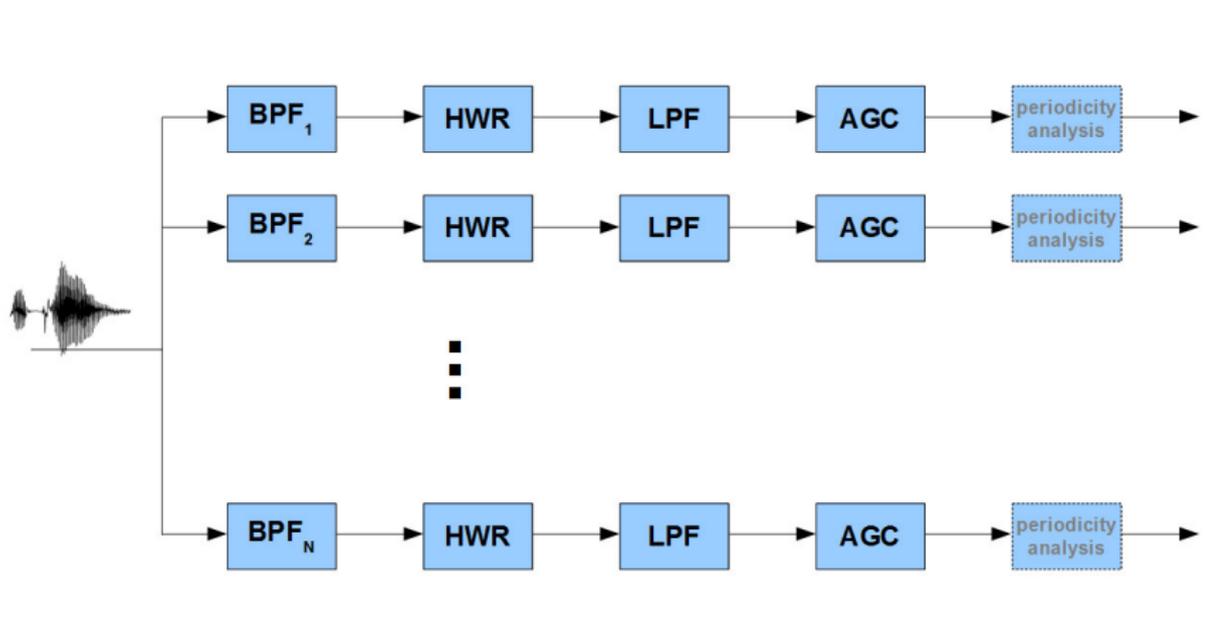
Current status of perceptually motivated algorithms

- ▶ The processing mechanisms of most of the steps are not exactly known
- ▶ Memory and adaptation not widely utilized
- ▶ Plenty of detailed knowledge about perception is ignored

Peripheral processing



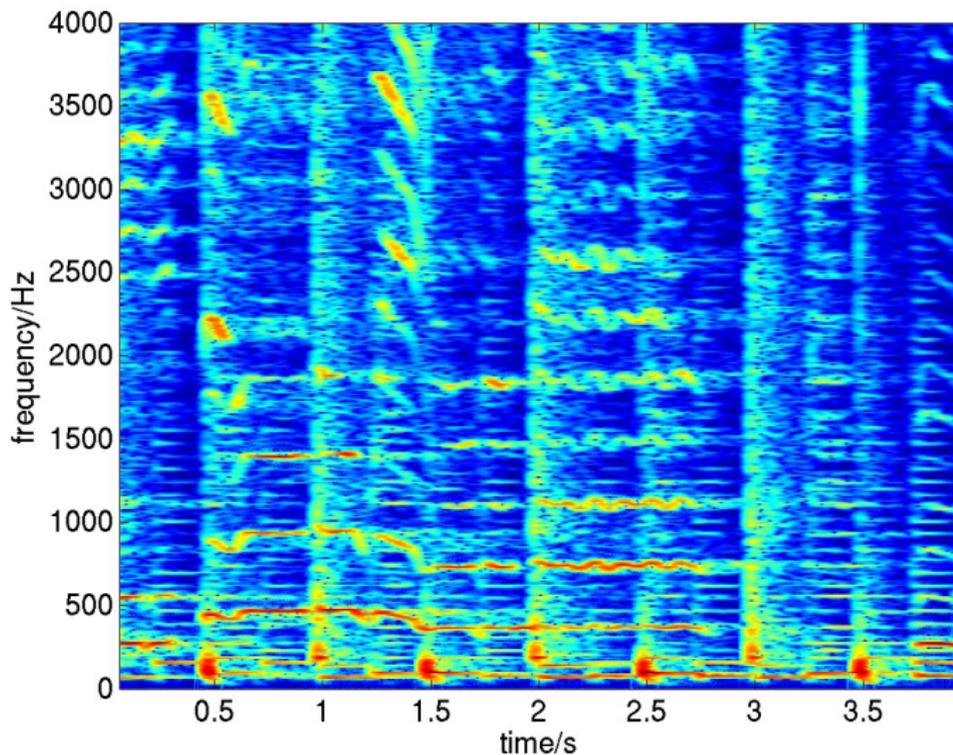
Standard model of peripheral processing in the auditory system



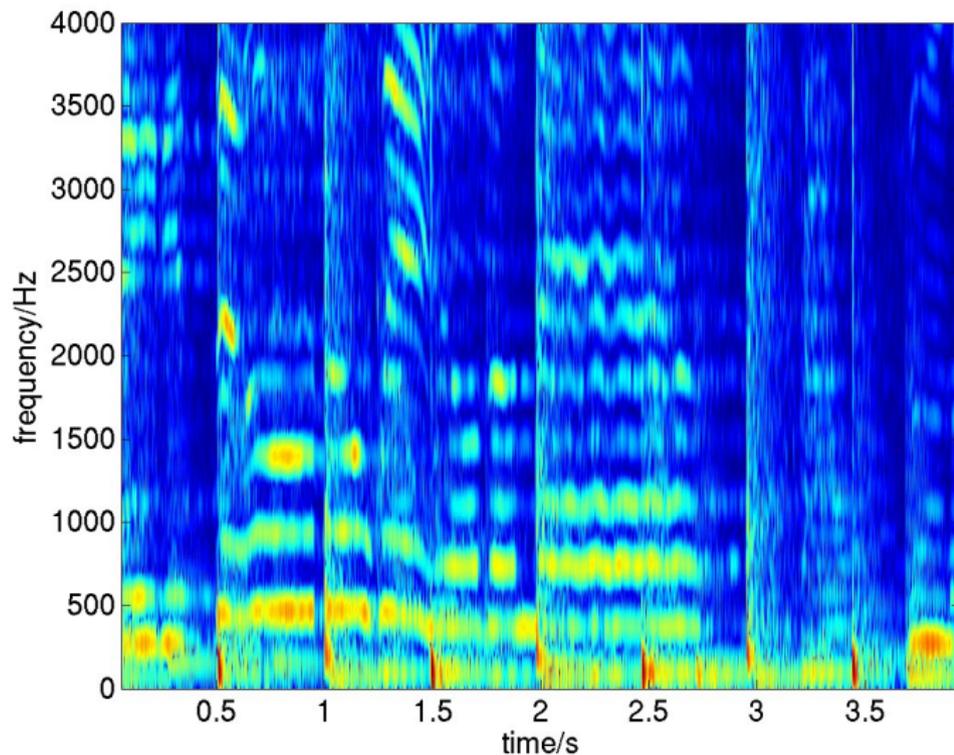
First stage: auditory filter bank

- ▶ The ear does *frequency analysis*

Spectrogram of music, calculated using a 120ms window



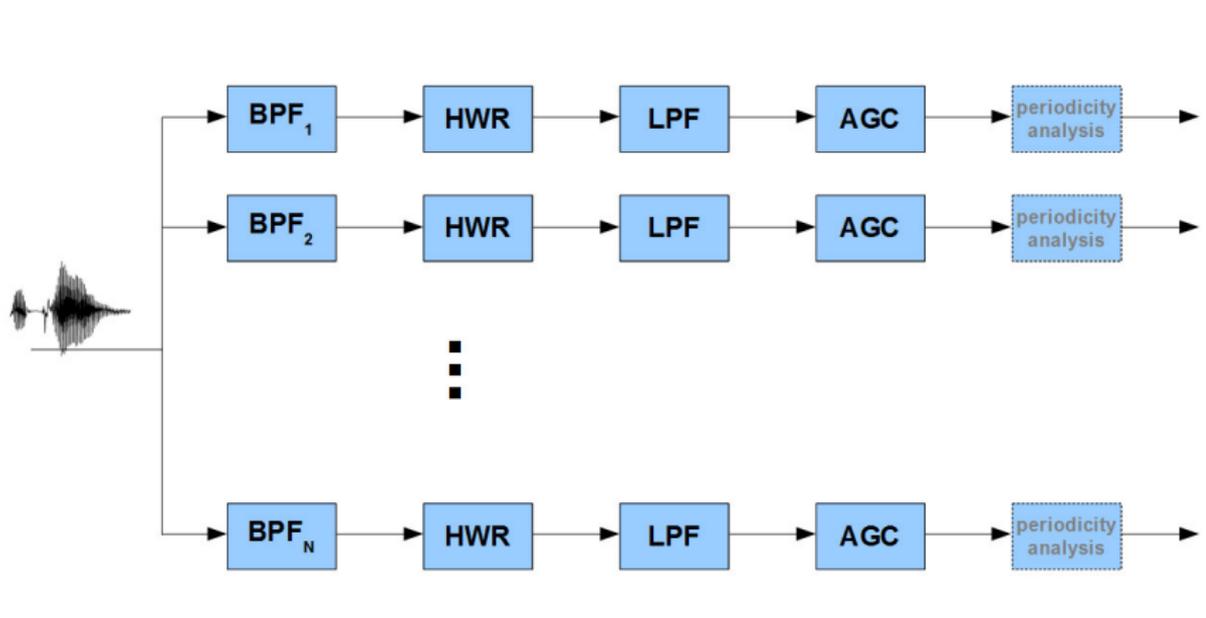
Spectrogram of music, calculated using a 10ms window



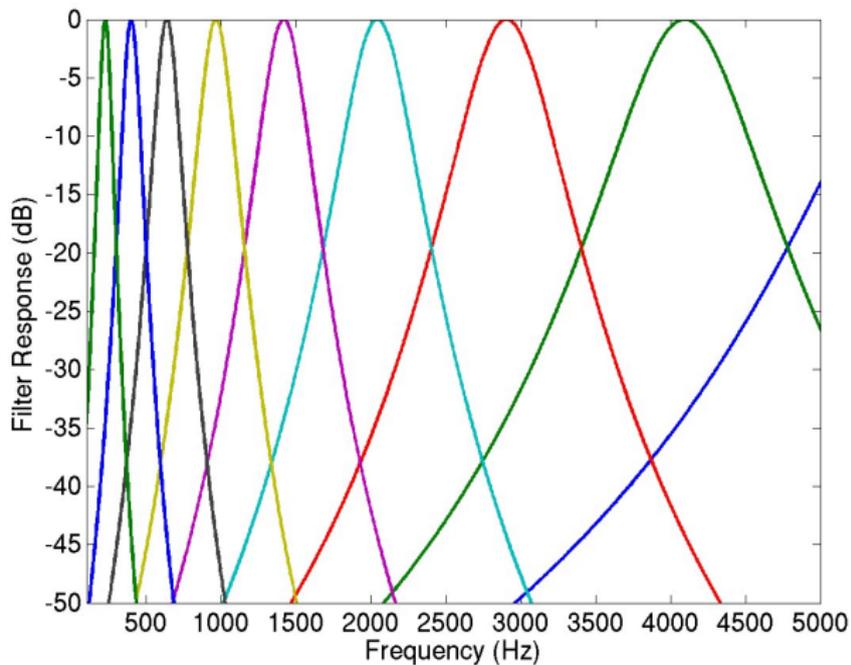
Time-frequency tradeoff

- ▶ Different tasks require different resolutions
- ▶ Tradeoff between the time and frequency resolutions
- ▶ Almost all audio analysis methods use a *fixed* analysis filter bank
- ▶ It has been shown that the auditory system extracts a set of highly *redundant* features
- ▶ This is completely different from standard audio features

Standard model of peripheral processing in the auditory system



Auditory filter responses



Effect of broad filter responses in feature extraction

- ▶ The responses of the auditory filters are rather broad
- ▶ Standard audio features calculate only the frame-wise energy within each auditory band
- ▶ Does not take into account the detailed structure of the signal within each band
- ▶ Based on the assumption of masking
- ▶ This approach is good in discarding irrelevant information in specific tasks — makes the signal invariant to e.g. pitch

Example: harmonic signal within three bands

Center frequency 300 Hz



Center frequency 900 Hz



Center frequency 2500 Hz



Envelopes of each band

Center frequency 300 Hz



Center frequency 900 Hz



Center frequency 2500 Hz



One band occupied by noise

Center frequency 300 Hz



Center frequency 900 Hz



Center frequency 2500 Hz

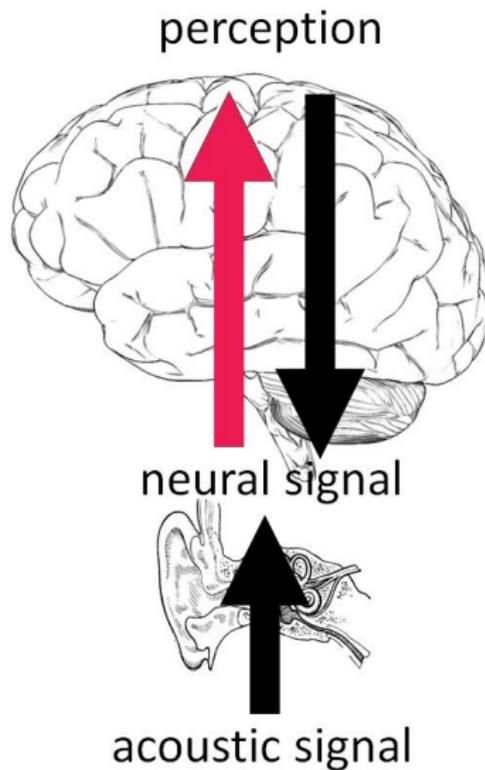


Retaining fine structure within bands

- ▶ The ear does preserve the fine structure of the signal — allows e.g. separating mismatched harmonics or noise
- ▶ When an auditory band is occupied by noise, using band energies as features does not allow distinguishing noisy channels

→ Invariance should be achieved by higher-level processing

Bottom-up processing



Bottom up processing

- ▶ Auditory grouping: elementary sound units are grouped to bigger entities
- ▶ Based on several cues:
 - ▶ Spectral proximity (closeness in time or frequency)
 - ▶ Harmonic concordance
 - ▶ Synchronous changes of the components: a) common onset, b) common offset, c) common amplitude modulation, d) common frequency modulation, e) equidirectional movement in spectrum
 - ▶ Spatial proximity.
- ▶ The exact grouping mechanisms are not known
- ▶ No proper knowledge about the interplay of different cues

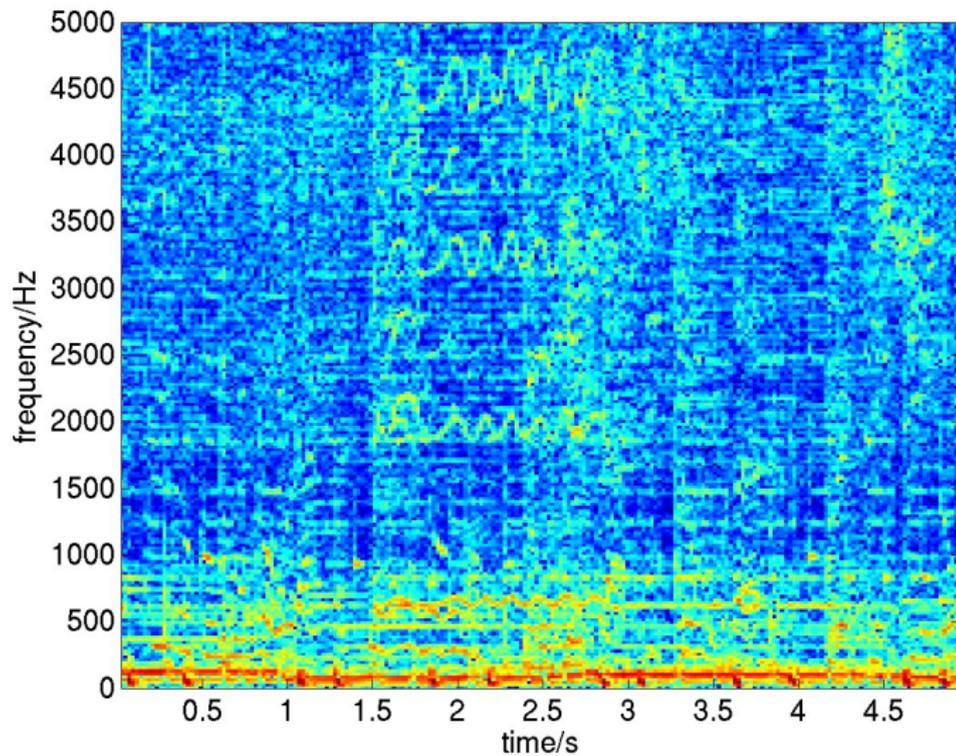
Computational bottom-up processing

- ▶ Clustering of elementary sound units
- ▶ Examples of elementary units: sinusoids, time-frequency cells, NMF components
- ▶ Existing grouping methods use typically only a small set of grouping cues (e.g. spatial or spectral)

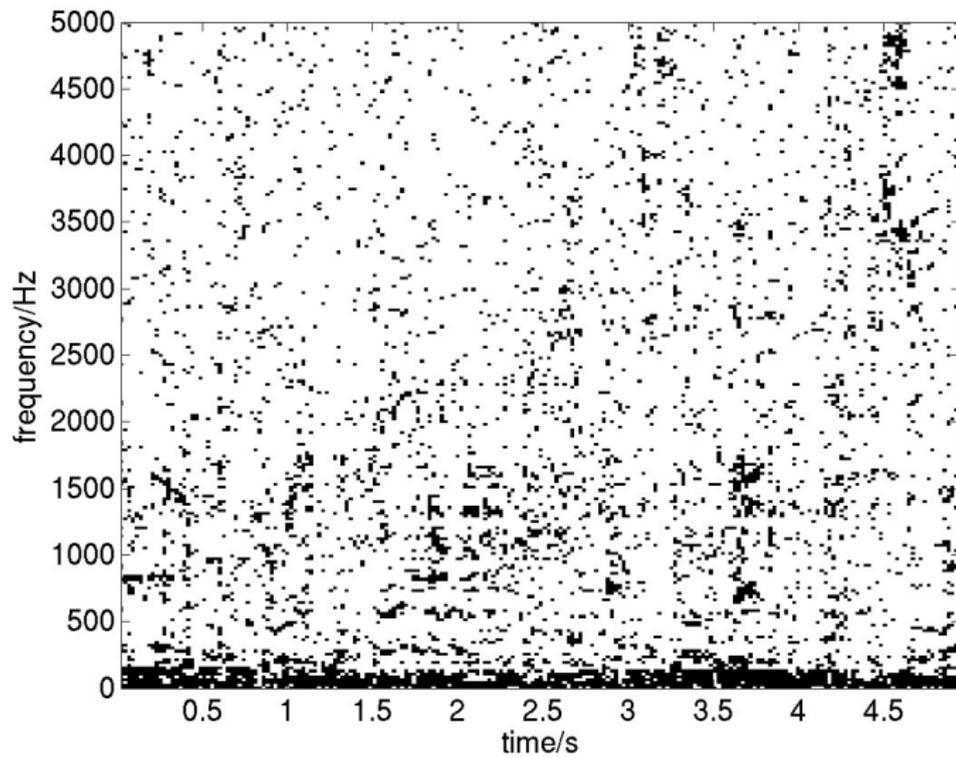
Clustering of time-frequency cells

- ▶ Commonly used in source separation
- ▶ Assumption: each time-frequency point of a mixture spectrogram belongs to only one source
- ▶ Motivated by masking (human's only perceive the dominant source in a frequency band) and sparseness (not likely that multiple sources have significant energy in the same T-F point)
- ▶ Estimate of time-frequency points of a source represented using a *binary mask*

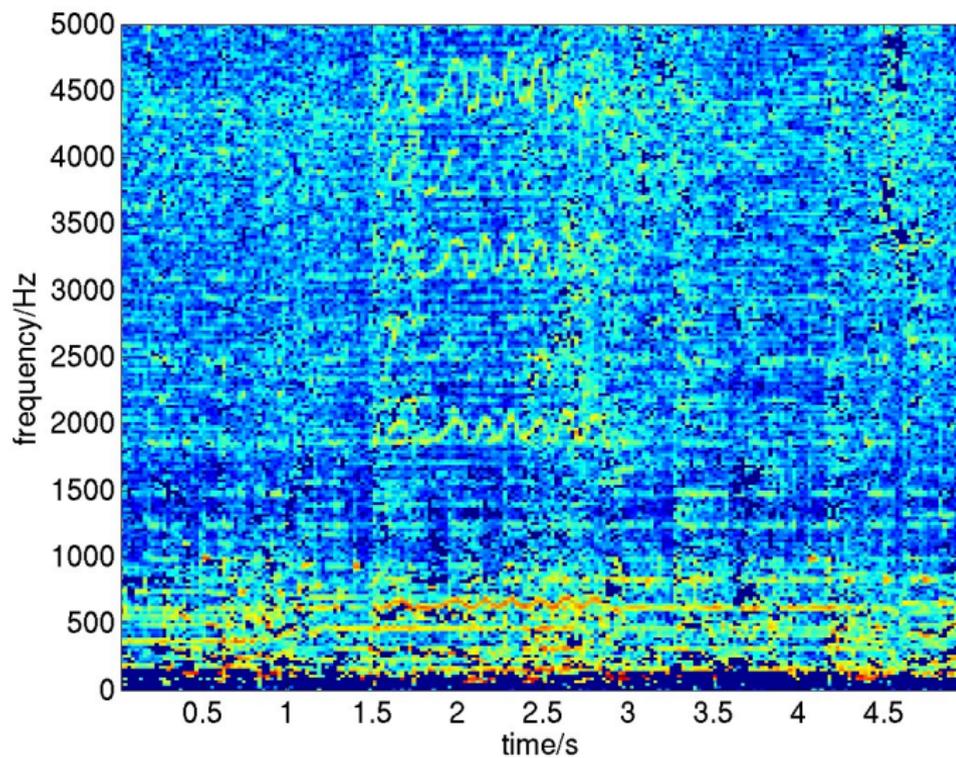
Example of a mixture spectrogram



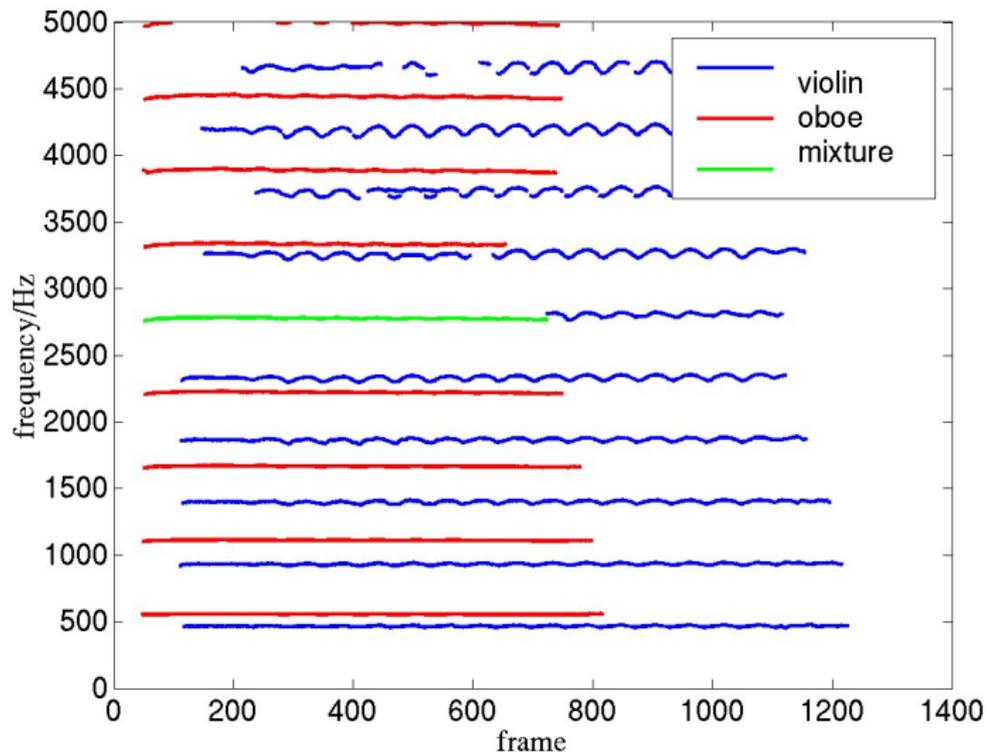
Estimated mask



Masked spectrogram



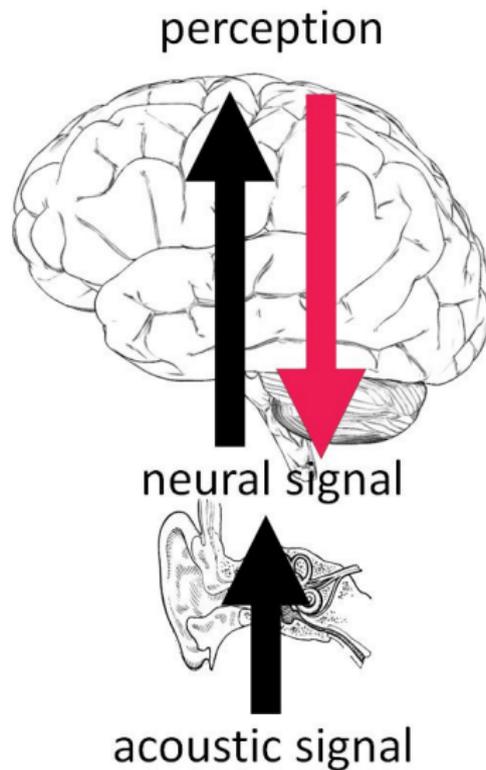
Clustering of sinusoids



Observations about computational clustering

- ▶ Possible to find simple/artificial cases where this works nicely
- ▶ In realistic conditions, clustering does not work
- ▶ Have to do some higher-level processing
- ▶ Example: speech fragment decoder (Barker et al.) which uses multiple hypotheses about fragments, and higher-level model for grouping

Top-down processing



Top-down processing

- ▶ Makes predictions and hypotheses about low-level representations, using higher-level information
- ▶ Schema-based processing: use of learned patterns
- ▶ The brain operates by doing pattern matching

High-layer patterns and low-level input

"cat"



"dog"



"cat"



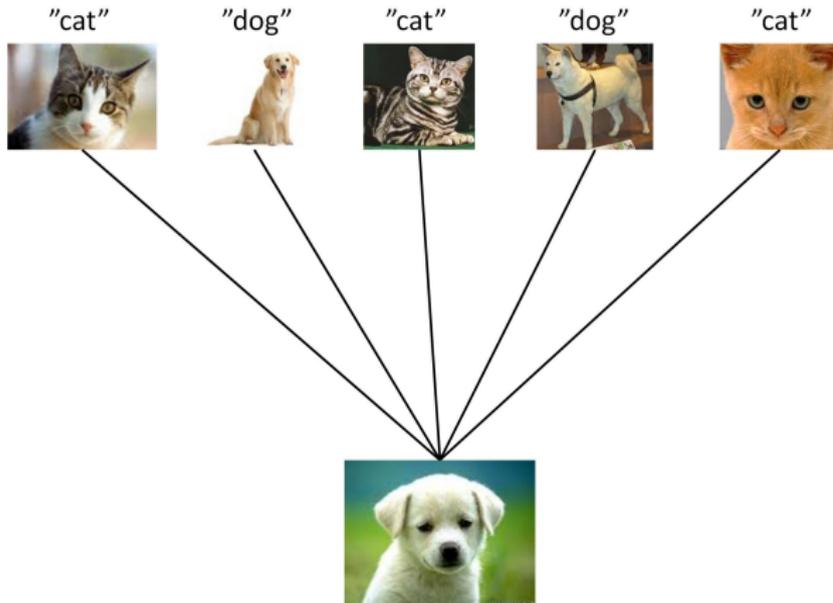
"dog"



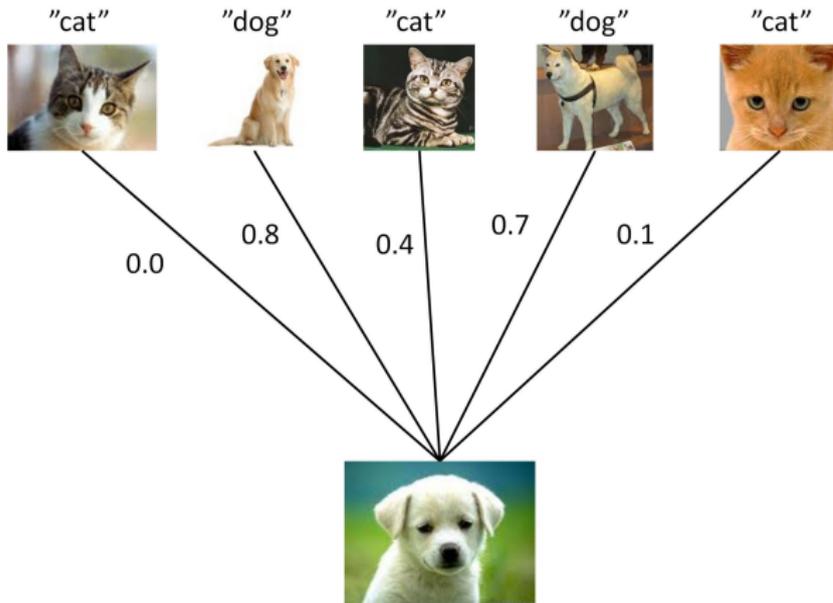
"cat"



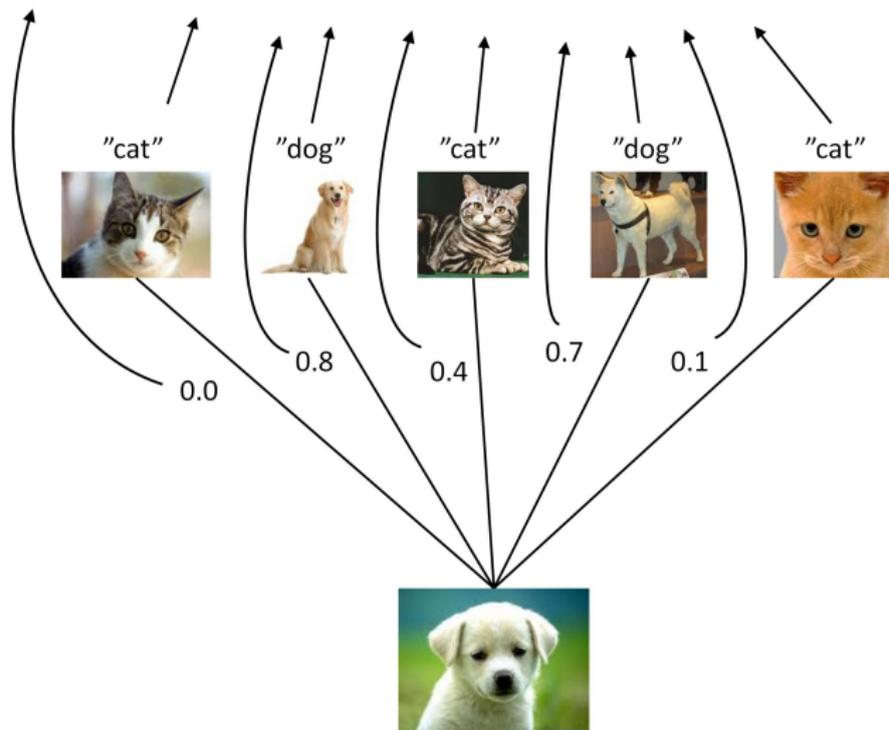
Pattern matching



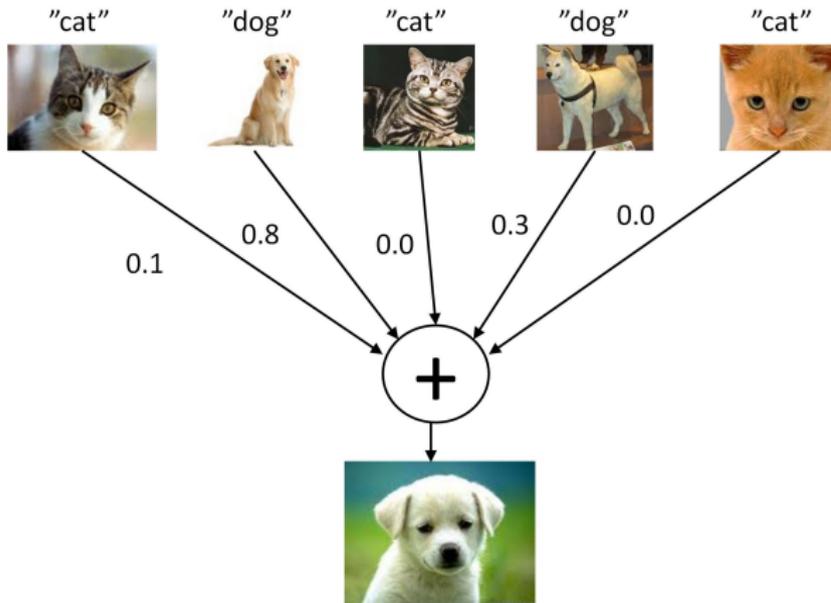
Similarity scores



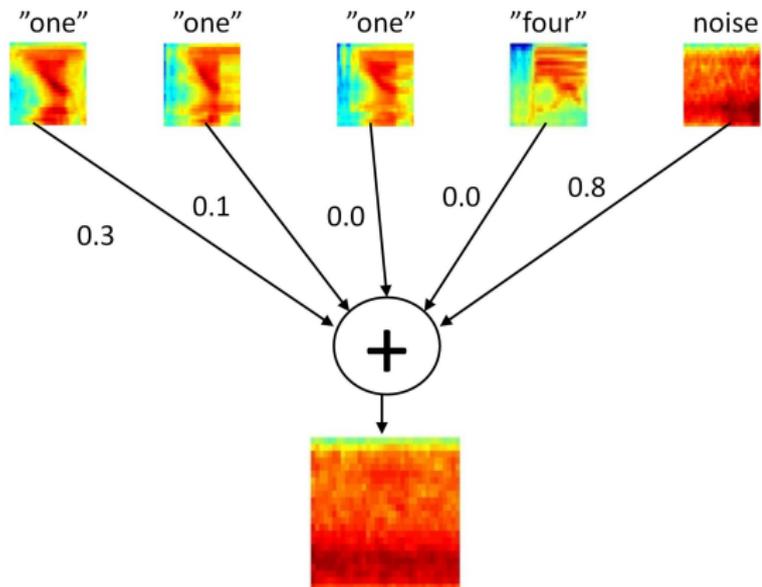
The scores and information about templates are used as features for higher processing levels



Sparse representation instead of similarities



Sparse representation for audio



Evaluation of results

- ▶ Using perceptually motivated methods may lead to *perceptually* good results
- ▶ On most audio analysis tasks, the performance is evaluated using *objective* measures
- ▶ Improvement of perceptual quality does not imply improvement of objective quality
- ▶ There is a need for perceptually motivated evaluation methods

Summary

- ▶ High-level auditory processing mechanisms could be utilized significantly better in computational algorithms
- ▶ Auditory perception also provides hints how to improve low-level processing methods
- ▶ Sparse representations are able to model many mechanisms of human perception