Real-world sound recognition

Tjeerd Andringa and Maria Niessen
{t.andringa,m.niessen}@ai.rug.nl

Auditory Cognition Group
Artificial Intelligence, *University* of Groningen

RuG

Introduction

Objective: develop novel approach to automatic sound recognition → recognize certain events in unconstrained input, i.e. any combination of sound in any environment

Why: standard sound recognition approaches function reliably only with input from a limited task domain in a specific environment

Inspiration from human perception → ecological theory of auditory event perception (Gaver, 1993; Gibson, 1979)
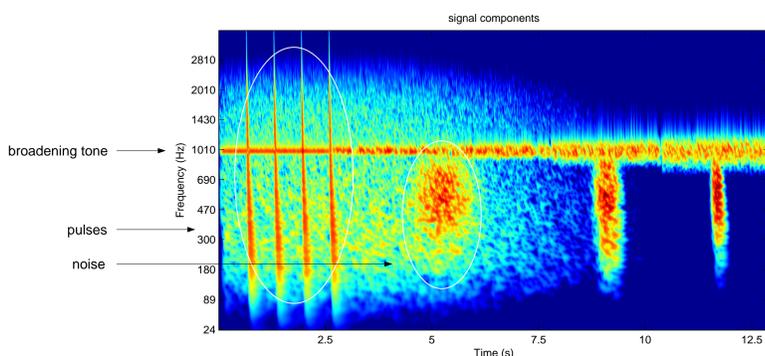
Event perception as a basis for automatic sound recognition:
• the sound signal is the result of the physical process which produces it
• the natural auditory system can recover events from the sound signal (Freed, 1990)
• we are interested in the event, not in signal details
• possibility to separate concurrent events
• limit solution space to physical plausible solutions: physical realizability

Paradigm to guide the development of the model:
(Andringa & Niessen, 2006)
• real-world sound recognition
• domain independence
• start from the natural exemplar
• physical optimality
• physical realizability
• limited local complexity
• testing with unconstrained input

features linked to physics of sound event → *signal components*:

Model for sound recognition

• process low level data to hypotheses about the sound event
• explicate semantically rich queries to expectations

Output: best hypothesis through the matching of bottom-up hypotheses and top-down expectations

Input: continuous unconstrained sound data, which is assigned to signal components, and signal component patterns are then processed to hypotheses about the sound event

The longer the system is processing data in an environment → the better its context model is able to generate hypotheses and the faster it will deal with bottom-up ambiguity.
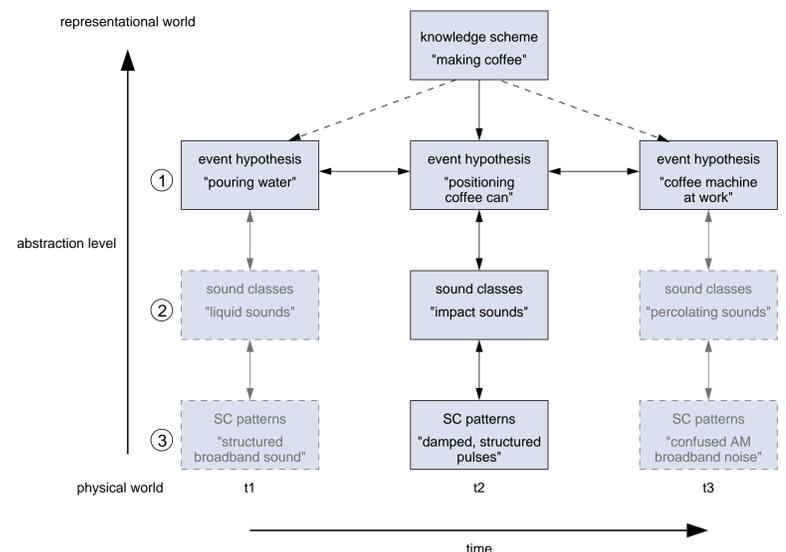
An example: making coffee

Lowest level description is coupled to signal components from the cochleogram:
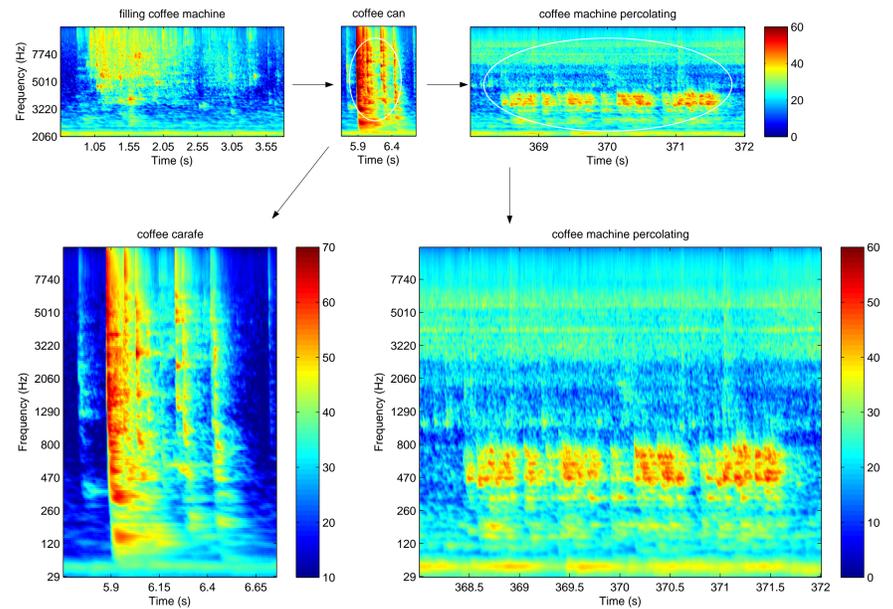
Structure: the positioning of the carafe in the machine (bottom left) results in a series of contacts between plastic and glass → pulses with internal structure, which do not repeat themselves and reduce gradually in intensity; some resonances, without harmonic pattern.

Automatic event recognition

• top level is highest abstraction level: most semantic content, minimal signal detail
• bottom level is lowest abstraction level: little semantic content, much signal detail
• succeeding best hypotheses (at t1 to t3), generated by low-level signal analysis, and matched to high level expectations
• more sound events are consistent with process → hypothesis is better supported → system will be more successful in predicting events
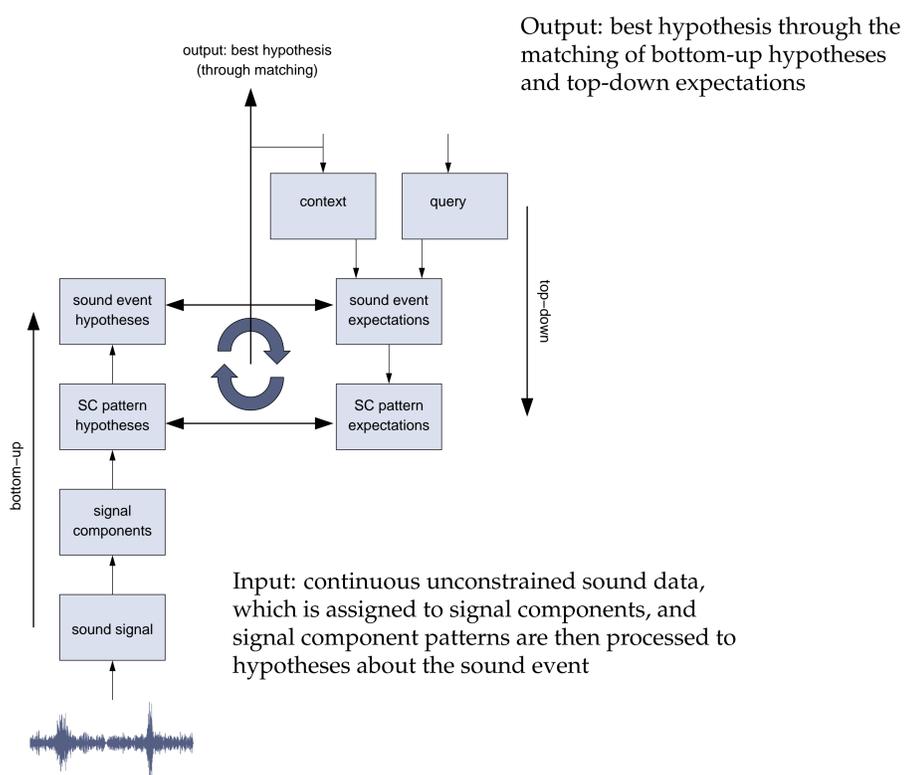

Example of recognizing a coffee making process at three levels:
1. sequence of activities common to making coffee
2. sequence of sound events to expect, associated with the activities
3. series of signal component combinations to expect, associated with the sound events

References

Andringa, T. C., & Niessen, M. E. (2006). Real-world sound recognition: A recipe. *LSAS*.

Freed, D. J. (1990). Auditory correlates of percieved mallet hardness for a set of recorded percussive sound events. *Journal of the Acoustical Society of America, 87*(1), 311-322.

Gaver, W. W. (1993). What in the world do we hear?: An ecological approach to auditory event perception. *Ecological Psychology, 5*(1), 1-29.

Gibson, J. J. (1979). *The ecological approach to visual perception.* Boston, MA: Hought Mifflin.