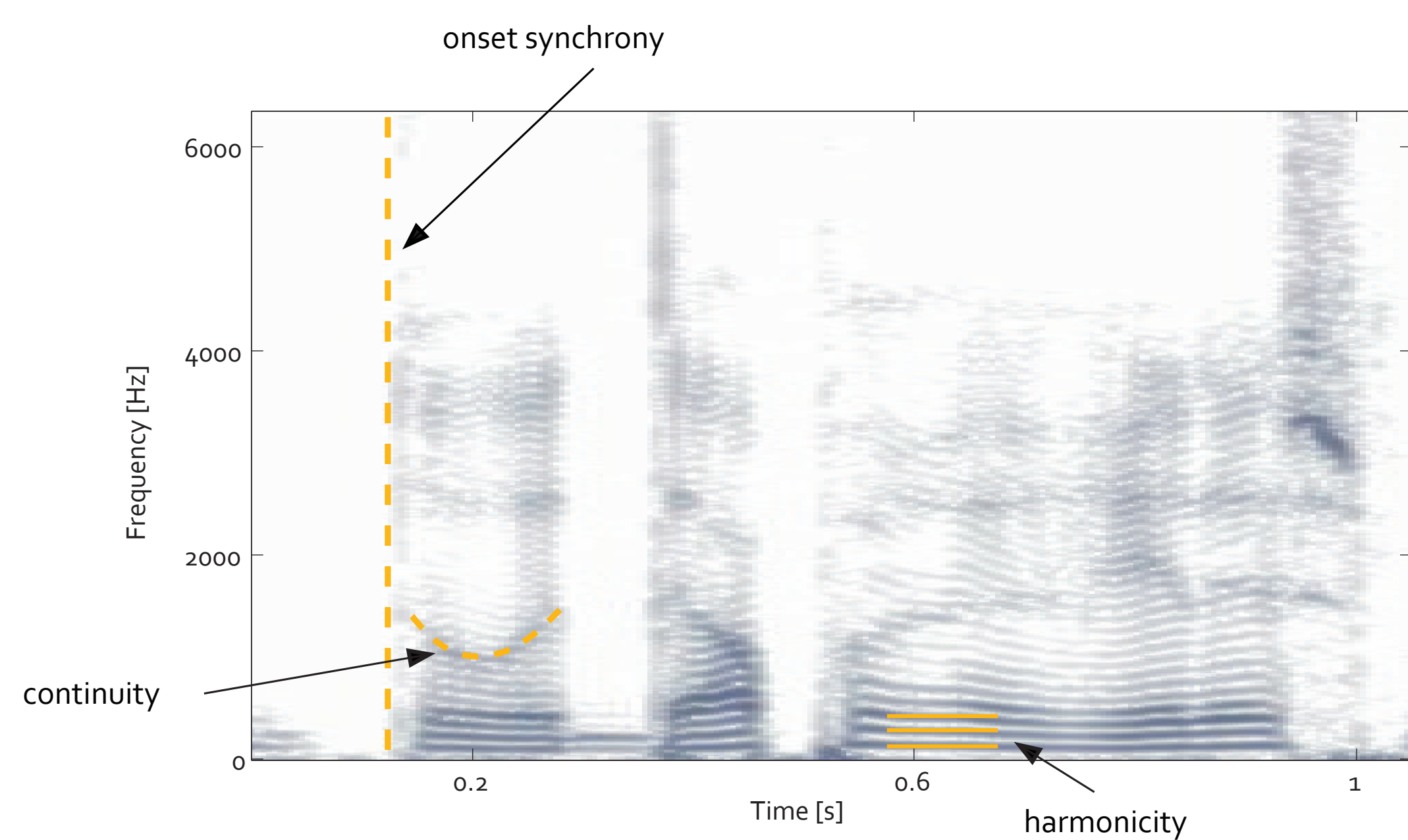


A computational model for auditory scene analysis

Maria Niessen, Ronald van Elburg, Dirkjan Krijnders, Tjeerd Andringa
Artificial Intelligence, University of Groningen

1 Introduction

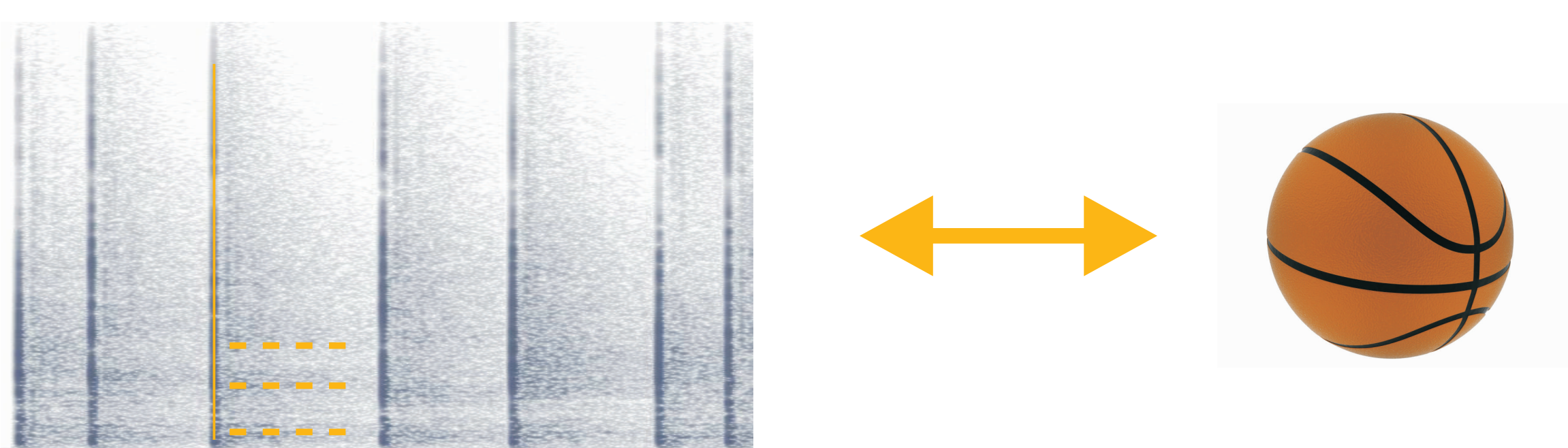
Primitive auditory scene analysis (ASA) is based on intrinsic properties of the acoustic environment. Acoustic elements such as continuity and proximity in time or frequency cause perceptual grouping of acoustic elements.



Various grouping attributes have been translated into successful signal processing techniques that may be used in source separation, e.g., to separate speech from background. However, separation is not enough to know:

What is the source of the sound?

A next step beyond primitive ASA is schema-based ASA, to give meaning to the source, i.e. to map bottom-up audio features to the meaningful content of an auditory scene.



3 Dynamic network model

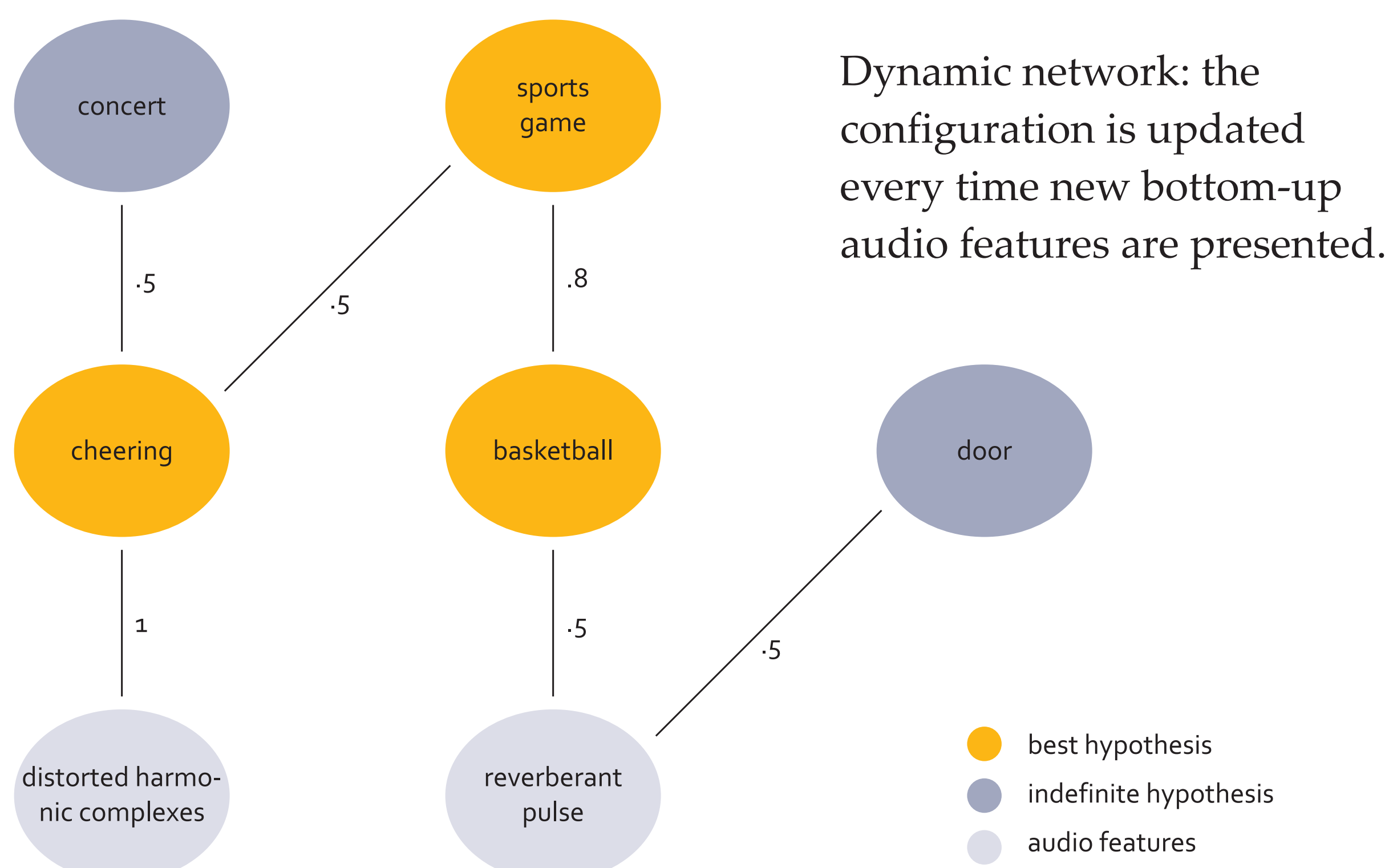
Bottom-up audio features are meaningless by themselves and require interpretation.

In complex real-world environments a sound signal may have different causes.

knowledge to give meaning to bottom-up audio features

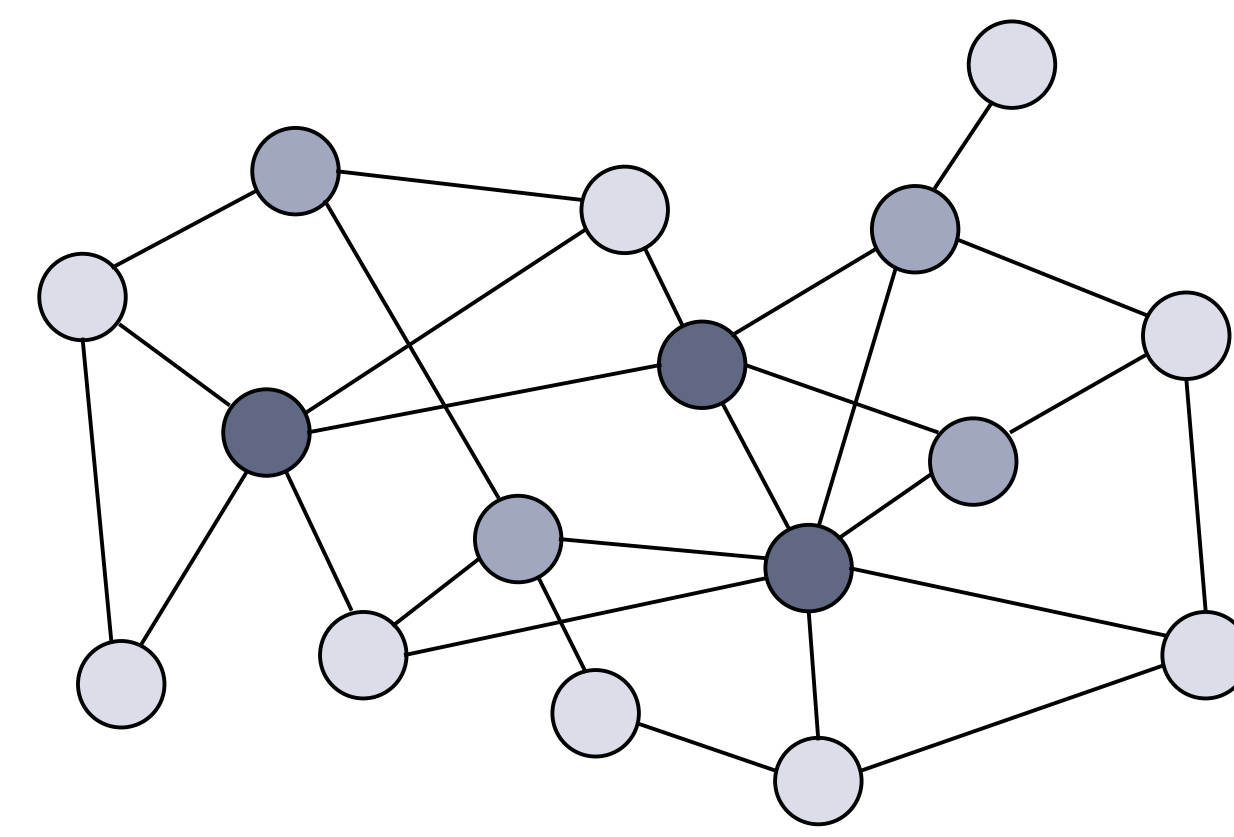
context to restrict the possible causes that the bottom-up features represent

Hypotheses based on bottom-up audio features are matched to expectations that are formed by knowledge of the relations between the events and the context.



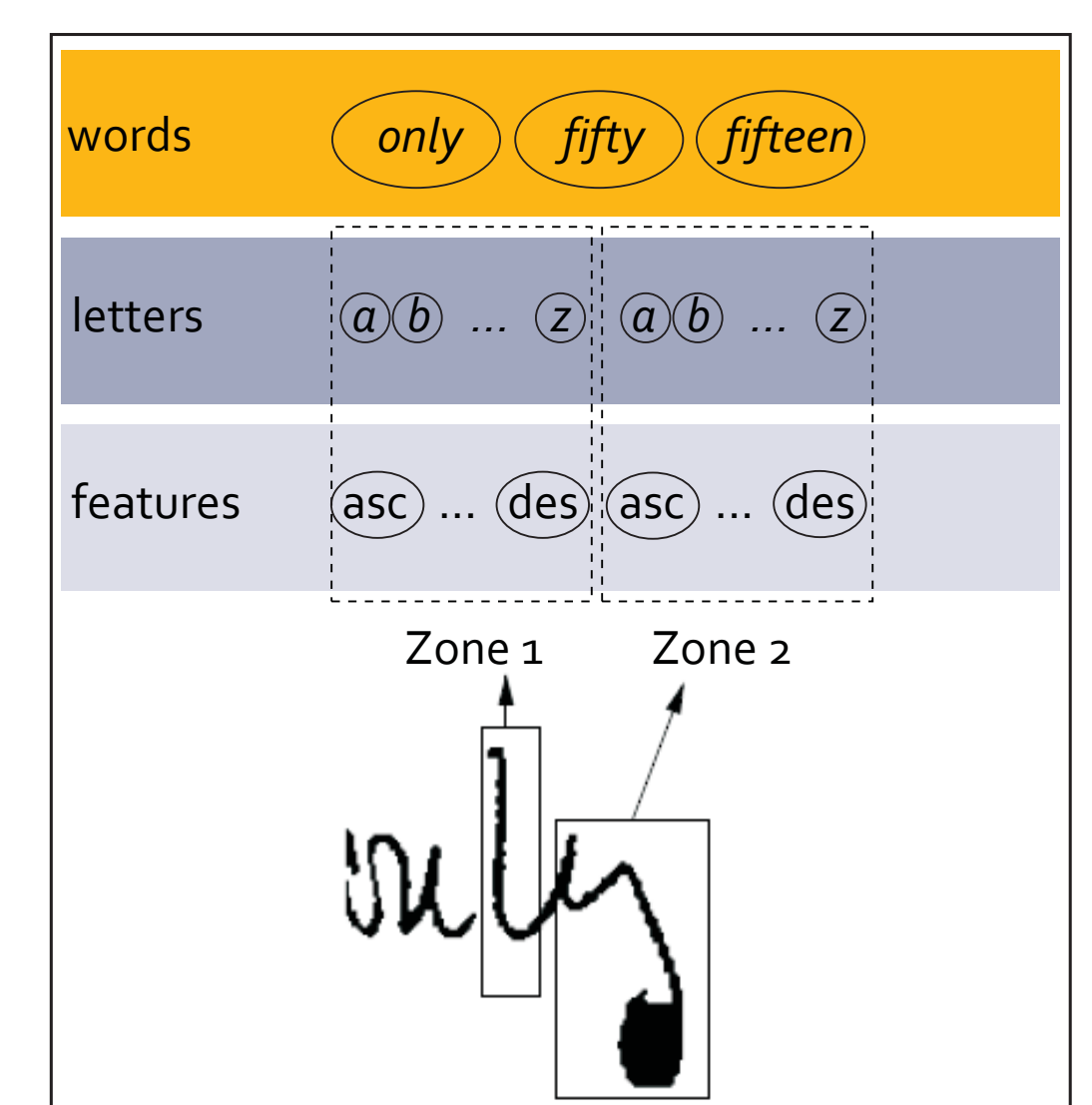
Network configuration for the identification of a reverberant impact sound in the context of cheering people. The best hypothesis at each level corresponds to a best explanation for the bottom-up evidence at that description level. The weights between the nodes reflect how strong the relations between events are.

2 Network models



Knowledge and context have been used in other research areas such as information retrieval. Often, this takes the form of a spreading activation semantic network, in which the nodes represent the states the network can be in, and the edges represent the prior probabilities that these states are encountered subsequently or together.

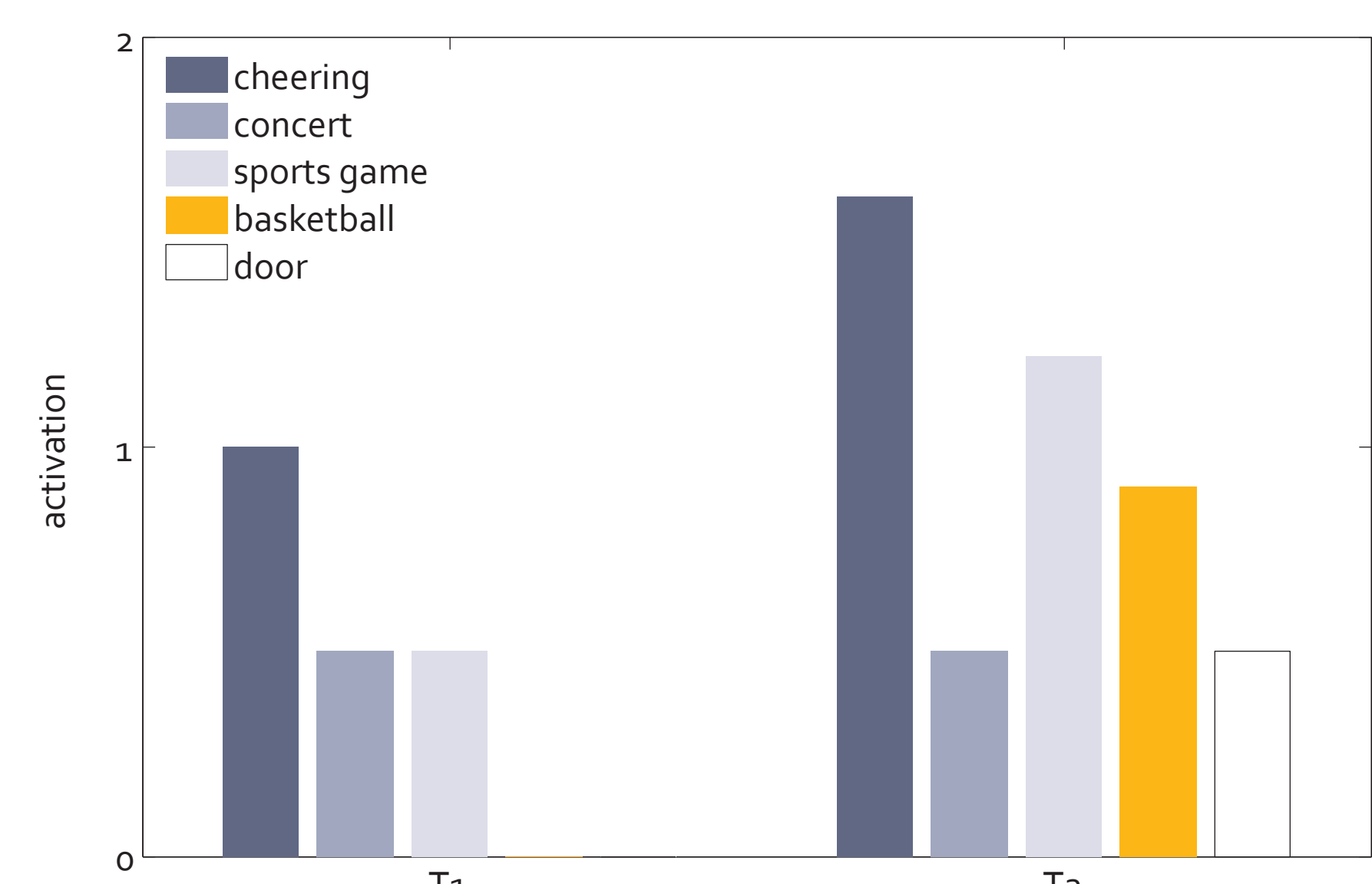
Instead of applying the spreading activation on static problems, e.g. handwriting recognition, we apply it to a dynamic domain: auditory scenes.



The use of context in automatic handwriting recognition, copied from Côté et al. (1998)

4 Spreading activation

All hypotheses hold a confidence value reflecting their support from relations to other events and the context. This confidence is computed by spreading activation through the network.



The activation (confidence value) of each hypothesis is a time-dependent weighed sum:

$$A_i = \sum_j w_{ij} A_j e^{-\Delta t/C}$$

j is a connected hypothesis at different updating times t
 w_{ij} is the weight of the relation between hypotheses i and j
 C is a constant decay parameter controlling the speed of decay
 Δt is the elapsed time since the hypothesis j stopped

5 Conclusions

We introduced a computational model for the analysis of dynamic auditory scenes. The differences with existing models of environmental sound recognition are

- the explicit use of knowledge
- the focus on identification of sound rather than classification

Our future work will include

- testing the model on databases of real events
- the appendage of grouped signal components as input to the model

More of our work can be found on www.ai.rug.nl/research/acg/