Further Challenges and the Road Ahead

J. Blauert, D. Kolossa, K. Obermayer and K. Adiloğlu

1 Introduction

Auditory modeling has traditionally been understood as a signal-processing task where the model output is derived from the acoustic input signals in a strict bottom-up manner by more or less complex signal-processing algorithms. The model output, then, consists of signal representations that are completely determined by the input signals. In other words, the output is *signal-driven*. It is then taken as a basis for predicting what is aurally perceived. This approach has also been taken for most of the application examples of auditory models reported in this volume [11]. However, notwithstanding the fact that the model output can predict actual aural percepts only in a very limited way, a further fundamental problem is left unsolved, namely, that human beings do not react on what they perceive, but rather on the grounds of what the percepts mean to them in their current action-specific, emotional and cognitive situation.

Inclusion of this aspect requires substantial amendments to the auditory models as they stand today. In addition to perception, assignment of meaning and formation of experience have to be dealt with among many other cognitive functions, for instance, quality judgements. In other words, models of binaural hearing have to be extended to models of binaural listening.

To this end, more advanced models will contain specific, interleaved combinations of signal-driven, that is, *bottom-up* processes and knowledge-based, hypothesis-driven, that is, *top-down* processes. The sub-cortical section of the auditory system has to be seen as an embedded component in a larger system where

J. Blauert (✉) · D. Kolossa
Institute of Communication Acoustics, Ruhr-Universität Bochum,
Bochum, Germany
e-mail: jens.blauert@rub.de

K. Obermayer · K. Adiloğlu
Neural Information Systems, Technische Universität Berlin,
Berlin, Germany

J. Blauert (ed.), *The Technology of Binaural Listening*, Modern Acoustics and Signal Processing, DOI: 10.1007/978-3-642-37762-4_18,
© Springer-Verlag Berlin Heidelberg 2013

signal-based representations are augmented by symbolic representations at different computational levels. Further, the models will contain explicit knowledge, accessible in a task-specific manner. Additionally, it will be necessary to replace current *static* paradigms of auditory modeling by considering the human being as an intelligent multi-modal agent that interactively explores the world and, in the course of this process, interprets percepts, collects knowledge and develops concepts accordingly. Consequently, models of binaural listening should incorporate means for the exploration of the environment by reflex- as well as cognition-controlled head-and-torso movements. Further, the role of input from other modalities—in particular, proprioceptive and visual input—has to be considered, since human beings are essentially multi-modal agents.

Figure 1 presents the architecture of a model that addresses the demands as described above. It is an architecture as currently discussed in AABBA [12]. The lower part of it schematically depicts the signal-processing, bottom-up processes and modules as can usually be found in today's models—see [37], this volume. The upper part represents modules that perform symbol processing rather than signal processing and are, to a considerable extent, hypothesis-driven, that is top-down controlled. In this upper part, various feedback paths are indicated, which are necessary for building a system that is capable of exploring and developing its world autonomously, thus gaining in knowledge and experience. In old-fashioned terms, one could actually call such a system a *cybernetic* one. It goes without saying that inherent knowledge at different levels of abstraction is required, particularly, when the system is supposed to perform quality judgment on auditory objects and auditory scenes that it has identified and analyzed [14].

2 A Framework for Cognitive Aural-Scene Analysis

Conceptualizing a framework for an artificial-listening system starts with the question of what the generic purposes of auditory systems are, or, in other words, why do humans listen at all? There is some consense in the field that three predominant reasons and, consequently, three modes of listening can be identified, namely,

1. Listening to gather up and process information from and about the environment, that is, to identify sound sources with respect to their nature and characteristics, including their positions and states of movement in space. This is also a prerequisit for appropriate action and reaction.
2. Listening for communication purposes. In many species interindividual communication is performed via the auditory pathway. In man, hearing is certainly the prominent social sense. It is, for example, much easier to educate the blind than the deaf.
3. Listening to modify one's own internal state, for instance, listening for pleasure, mood control, cognitive interest, and so on.

The different listening modes determine the strategy that listener persue in given situations and, thus, draw heavily on their cognitive capabilities. This is another case for including cognition in a framework of auditory listening, although it is still not clear as to what extent these modes of operation can be mapped onto a unified architecture. However, it is conceivable that at least for the symbol-processing part in an artificial-listening system some universal strategies can be employed—at least up to the point where meaning is converted into action, or where internal states need to be represented and changed. For the framework introduced in the following, the goal of understanding the external auditory environment, that is, item (1.), provides the main focus. Yet, the authors are aware of the other modes and shall try to accomplish them as well in the further course of the model development.

In the following, the overall function of the model is described in general terms in accordance with the framework depicted in Fig. 1.

The acoustic input to the model is provided at block (a) by a replica of a human head with two realistically formed external ears and two built-in-microphones. This artificial head is connected to a shoulder piece to form, together with the head, a head-and-torso simulator. The head is capable of three-degrees-of-freedom movements with respect to the shoulder piece, namely, rotating, tipping and pivoting. The head-and-shoulder simulator is mounted on a movable cart, which allows for further two degrees of freedom for translatory movements. There are sensors to monitor the positions of head, shoulder piece and cart with respect to each other and to an external reference point. The movements are enabled by actuators which can be remotely controlled. Which of the possible sensors and actuators are actually implemented, depends on the specific tasks that the model is specified for. Depending on the respective tasks, the equipment may further be fitted with sensors for additional sensory modalities, such as visual or tactile sensors.

The audio signals from the two microphones are fed into block (b), which represents major functions as are regarded relevant to be implemented by the human subcortical system up to the midbrain level. The components of this block account for functions that are attributed, for example, to the *middle ears* and the *cochleae*, to the *superior olivary complex*, SO—including *medial superior olive*, MSO, and *lateral superior olive*, LSO—or to the *inferior colliculus*, IC. Those functions as well as their computational implementations are described in more detail in [37], this volume.

The output of block (b) is represented within the computational model through a multidimensional, binaural-activity map including, for example, the dimensions intensity of activity, frequency, lateral position or time—see block (c). This kind of representation is inspired by the existence of activity maps at the midbrain level of animals for coding acoustic features, such as spatial locations of sound sources, fundamental frequencies and spectra and/or envelope characteristics of the acoustic source signals. A specific example of such a computational binaural-activity map is depicted in Fig. 1, namely, a map depicting binaural activity as generated by the binaural impulse responses of a concert hall.

The next step in the model, block (d), has the task of identifying perceptually relevant cues in the activity maps and, based on these cues, organize the activity

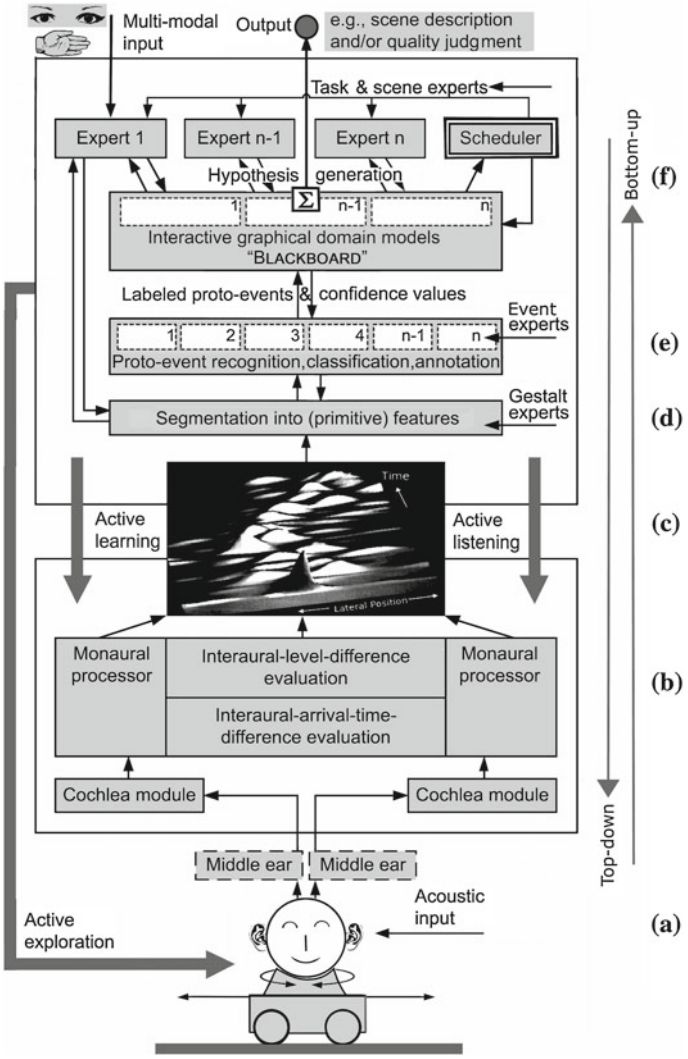


Fig. 1 Schematic of the architecture of a comprehensive model of binaural listening. **a** Head-and-torso simulator on a mobile platform. **b** Signal processing in the lower auditory system. **c** Internal representation of binaural activity. **d** Rule- and/or data-driven identification and annotation of perceptually-salient (primitive) features. **e** Rule- and/or data-driven recognition, classification and labeling of proto-events. **f** Scene-and-task representation, knowledge-based hypothesis generation, assessment and decision taking, assignment of meaning

maps into segments that represent specific primitive perceptual features. The cues can be signal attributes in the temporal or spectral domain, such as autocorrelation, decrease and centroid, effective duration, energy, interaural arrival-time differences, interaural cross-correlation, interaural-level differences, log-attack time,

modulation amplitude, modulation frequency, roll-off frequency, spectral centroid, temporal increase, time frame, zero-crossing rate—just to name a few. Yet, they can also be estimators for *"sensations"* such as pitch, timbre or coloration, loudness, sharpness, roughness, spaciousness, and reverberance, as can be calculated with specialized signal-processing algorithms. The features rendered are typically based on an ensemble of different cues, whereby primitive schemata like the *Gestalt rules* [6, 17] are considered in the formation process. The rendering process can be rule-driven, that is, exploiting prior knowledge, or data-driven, that is, based on statistical procedures—see Sect. 2.1 and 3 for more details of relevant machine learning methods. Multiple processes of this kind may act in parallel, also as a combination of data-driven and rule-driven approaches.

The primitive features derived from the activity map provide the input to the next model stage, block (e), where sets of features are interpreted as indicators for specific auditory events. These indicators are called *proto-events* here, since their actual character is subject to statistical uncertainty. The formation process for proto-events may, for instance, follow a sequence of detection, classification and annotation. Again, rule-driven as well as data-driven approaches may be used. Clearly, the successful extraction of proto-events depends on whether appropriate feature sets have been chosen in the beginning. In the context of an artificial system, feature-selection techniques can be employed, and small sets of informative features can in principle be learned for given combinations of (classes of) auditory tasks and (classes of) auditory environments. However, in the context of bio-inspired processing as for human-listening modeling, this becomes a highly non-trivial task. One possible approach to proceed is to conduct model-driven psychoacoustic experiments and to ask, whether human listeners employ the same features as the artificial system suggest.

By the way, blocks (d) and (e) have been described here as two sequential processing steps. However, for certain statistical procedures, such as being used, for example, in the machine-learning field, the difference between primitive features and proto-events may not always be clear-cut and processing steps (d) and (e) may well be combined into one. It is at this model stage, that a *transition from signal processing to symbol processing* takes place, since the proto-events can be represented by symbols.

The last stage, block (f), represents the world-knowledge of the modeling system—among other functions. At this stage, contextual information is used to build task-related representations of the auditory scene, namely, prior knowledge is integrated, hypotheses about auditory events are generated and validated, and meaning is assigned. The auditory scene is evaluated, decisions are made and signals or commands may be sent back down to the lower processing levels—blocks (a) and (e).

In the framework of Fig. 1, a so-called *blackboard structure* [20, 27, 28] is proposed for this purpose. It works as follows. The input from the lower stages is put on a "blackboard", which is visible to a number of specialized experts, namely, computer programs that try to interpret the entries on the blackboard based on their respective expert knowledge. There can be various different experts, for instance, acoustic experts, psychoacoustic experts, psychologic experts, experts in spatial hearing,

experts in cross-modal integration of vision, tactility, proprioception etc., speech-communication experts, music experts, semiotic experts, and so on—depending on the specific task that a listening model is constructed for.

Once an expert finds a reasonable explanation of what is shown on the blackboard, it puts this up as a hypothesis to be tested against the available entries. The hypothesis will then be accepted or rejected based on rules or on statistical grounds. For the control of the activities of the experts, a special program module, the *scheduler*, is provided.

The scheduler acts like the chairman of a meeting. Firstly, it determines the order in which the individual experts intervene, controls the statistical testing and makes a decision regarding the final outcome—which may well be a mixture of various accepted hypotheses—compare [30] as to how to provide such a mixture. Secondly, it will also select groups of experts as well as modify the computations performed by them on the bottom-up data according to the current goals, such as extracting the *"what"* versus the *"where"* of a sound event.

In the light of the main focus of the current chapter, that is, the analysis and assessment of aural objects and scenes, two types of models are of particular relevance, namely, (i) *object models*, to understand single aural perceptual entities known as aural objects, and (ii) *scene models*, to understand interactions between aural objects in arrays of objects, for instance to cover questions like: Which of the objects may be simultaneously aurally present. Hereby, the following definitions may apply.

Objects are perceptual entities that are characterized by specific attributes and invariances and by their relations to other objects

Scenes are arrays composed of multiple objects, again specified by specific attributes and invariances

For creating models of aural objects and models of aural scenes, a wide range of knowledge can be exploited, such as rules of how object and scenes develop, physical knowledge of type, location, and movement of sound sources, perceptual knowledge such as Gestalt rules, or semantic knowledge with respect to the information that the sound sources may intend to convey. Here, so called *graphical models* are proposed to represent these models and, at the same time, act as the blackboard. *Graphical domain models* allow for invention both from block (e), that is the proto-event level, as well as from the experts' level. For more details see Sect. 2.2.[1]

The outcome from block (f) and, thus, the output of the listening model at large, may be a description of an aural object or the description of an aural scene. Yet, it is further planned to develop the model to such a state, where it can assess aural objects and scenes with regard to their perceptual quality. Judgment on quality requires a further processing stage for the following reason. Quality, in general terms is the amount

[1] Graphical models are convenient when it comes to the implementation of a working artificial-listening system, but whether and—if yes—how it actually maps to the processes which integrate and disambiguate sensory information in the human brain remains a matter of future research. It has been suggested that neural systems implement *Bayesian inference* including even *belief propagation* [24, 25], but there is also evidence, that competition between *neural assemblies* and an *attractor dynamics* [23] may play an important role in sensory processing.

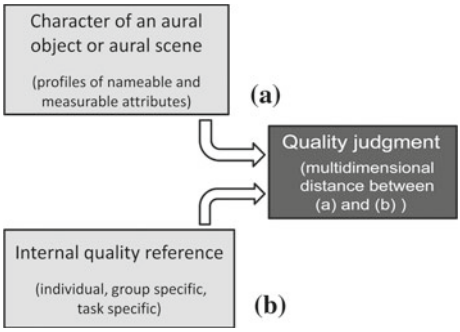


Fig. 2 Quality judgment as a multi-dimensional distance measure between the character, (a), of an item and a reference, (b), representing expectations regarding this item

as to which an item fulfills expectations with regard to it [26]. A quality judgment thus requires a set of measured and nameable features of the item under consideration, that is, its *character, (a),* and a set of values expected for these features, the *reference, (b)*. A quality judgment can then be seen as a multi-dimensional distance measure between those two—see Fig. 2. It follows that the expert must have internal references to apply when judging on quality. These reference are individual and/or group dependent and task specific [10, 13]. To construct them can be impossible, where internal quality references are concerned, and it is tedious even in the best of cases.

To avoid this, it is also possible to perform the quality assignment as the result of machine learning. In this case, the machine-learning algorithm needs the quality judgments assigned to specific situations as a-priori information. Whether this turns out to be less tedious in the end than collecting information about internal references, remains to be explored for each specific case. Also, it is of advantage for many applications to have direct and explicit access to the internal references behind the quality judgments. With machine learning this would require further analysis.

At this point, some remarks on *signal-driven*—bottom-up—and *hypothesis-driven*—top-down—processing are due, since they proceed in an interleaved way in the model framework as proposed here. In purely signal-driven processes the output is completely determined by the input. If the processing requires multiple variables to be combined, with each of them having a number of possible states, this can quickly lead to an immense number of potential output states—combinatorial explosion—which all have to be followed up and evaluated until a final decision has been reached. In top-down processing, in contrast, the number of states to be evaluated is substantially reduced, as the process *knows what to look for*, that is, focuses attention on states which *make sense* in a given specific situation. Of course, such a strategy is limited to known scenarios, unless means for adaption and assimilation are provided.

To avoid this deadlock and for other reasons, Fig. 1 provides various feedback paths, some more specific, others more general. The general ones originate from the concept that the listener model, that is, the *"artificial listener"*, actively explores its aural world and thereby differentiates and develops it further in an autonomous way, very much like human beings do. Following this line of thinking, it is attempted to

model listeners according to the *autonomous-agent* paradigm, where agents *actively learn* and *actively listen*. Since the listener model can deliberately move its sensors about in the space to be explored, it can use proprioceptive cues besides aural ones to perform these tasks. Cross-modal cues, like visual and/or tactile ones [4, 38] may be included if appropriate. Active learning and active listening are further discussed in Sect. 3.

2.1 Feature Extraction and Proto-Event Detection

In the framework of Fig. 1, statistical machine-learning techniques are planned for the extraction of primitive features, block (d), and for the formation of proto-events, block (e). As their input, these techniques require signal partitions and/or features that carry information that is relevant for what is finally perceived and considered to be meaningful.

By using such input, machine-learning models can be developed that are able to extract proto-events from autonomous and/or interactive environments. These proto-events can then be used at higher levels for comparison and verification—for example, in graphical models, see Sect. 2.2—as well as for providing feedback to lower-level model stages. The final goal is to arrive at proto-events that make sense to human beings, particularly in the light of previous experience.

In the machine-learning field, agents learn tasks from data that are provided to them by the environment, that is, tasks are learned by induction. Within the processing stages of blocks (d) and (e), in other word, agents will primarily perform pattern recognition tasks. Following the machine-learning paradigm, agents will first have to undergo a learning phase during which informative acoustic features are selected for input, and the agents' parameters are tuned in such a way that the pattern-recognition task can be performed sufficiently well. In the following recognition phase, these agents will then fulfill their "duty", which is to extract the relevant acoustic features from the input signal and to combine them for the detection, classification and annotation of proto-events.

Traditionally, one distinguishes three learning schemes on the basis of what kind of information is available to the agent. In *supervised* learning, the agent is provided with the acoustic input simultaneously with the correct annotation. In *reinforcement* learning, the agent is provided with the acoustic input, but the environment provides only a summary feedback signal that tells the agent whether the annotation was correct or not. In *unsupervised* learning, finally, only the acoustic input signals are available, and the agents' task is to utilize statistical regularities within the acoustic environment in order to generate a new representation that is optimal—given some predefined learning criteria.

Although reinforcement learning is a key learning scheme when it comes to human beings, its machine-learning counterparts are computationally expensive and require extremely long learning periods. Therefore, it is currently not advisable to use it in the framework of Fig. 1. As to the other two paradigms, unsupervised-learning

paradigms are better suited for learning feature representations—block (d)—while supervised-learning paradigms are generally better suited for generating symbolic representations—block (e).

Successful learning and recognition strongly depends on the representation of auditory signals and scenes. While many different approaches should be implemented, evaluated and compared, a particularly interesting class of representations are the biologically-inspired spike-based representations [48]. These representations are typically sparse and provide a decomposition of a complex auditory object into a pattern of brief *atomic events*. This decomposition can, for example, be derived from the binaural-activity map generated at block (c), where the atomic events would then be localized in both auditory space and time. Another example is discussed in more detail in the next subsection.

Although many auditory objects are distinctly localized in time and space, they may still occur at variant spatial locations and points in time. Vector-based representations typically have difficulties capturing this and other types of variability, for instance, different durations and/or spectral shifts. Relational representations, where auditory events are described by their similarity to other auditory events rather than by their individual features, are much better suited, because the required invariances can often be built into the similarity measure in a straightforward way. In addition, many kernel-based machine-learning methods have been devised over the last 20 years that naturally operate on relational representations, including, for example, the well known support-vector machines. Although standard kernel methods impose certain constraints on the similarity measure, that is, the *kernel*, extensions have been suggested, such as by [34, 35], that can also be applied to a wide class of similarity measures for spike-based representations. Two examples for the supervised and unsupervised learning paradigms are described in the according subsection herewithin.

Sparse Event-Based Representation

Spike-based representations [48] provide a decomposition of a given sound signal via a linear combination of normalized basis functions taken from a predefined or learned dictionary. This kind of representation will now be illustrated by a simple example from the auditory domain. Let $x(t)$ be a monaural sound signal and let $\gamma_{f_k,t_k}(t)$ be the basis functions, then

$$x(t) = \sum_{k=1}^{K} a_k \gamma_{f_k,t_k}(t) + \epsilon_{K+1}(t). \tag{1}$$

Every basis function, $\gamma_{f_k,t_k}(t)$, in (1) corresponds to one atomic event located at time, t_k, in the auditory stream. The corresponding coefficient, a_k, and "property", f_k, characterize this event. The parameter f_k determines the type of basis function taken from the dictionary and can, for example, be the center frequency of a time-frequency

localized filter from a given filterbank. a_k then describes how well the filter function locally matches the auditory stream. The most efficient representation of the type defined by (1) is one that achieves a small residual, $\epsilon_{K+1}(t)$, for a small number, K, of atomic events. A greedy way of iteratively constructing such a representation employs the matching-pursuit algorithm [40]. During each iteration, k, a basis function, γ_{f_k,t_k}, is selected that maximally correlates with the residual signal, ϵ_k, remaining from this iteration, that is,

$$(f_t, t_k) = \text{argmax}_{f_m,t^*} < \epsilon_k(t), \gamma_{f_m,t^*}(t) > . \tag{2}$$

The total number, K, of iterations determines the number of events with which a particular auditory object is described, in other words, the level of sparseness as well as the accuracy of this representation, namely, the magnitude of the remaining residual. There is a trade-off between both. However, since the goal of this representation is *not* to reconstruct the sound at later processing stages, the absolute size of the residual is not so important, and the focus should be on creating a representation that uses a small number of basis functions, that are most informative for later classification and annotation. Different filter functions are suitable for generating an overcomplete dictionary—such as Gabor atoms, Gabor chirps, cosine atoms, or Gammatone filters. Gammatone filters are popular for auditory-adequate filtering, because a subset of them approximates the magnitude characteristics of human auditory filters [43]. Figure 3 shows the spike-based representation for a specific sound

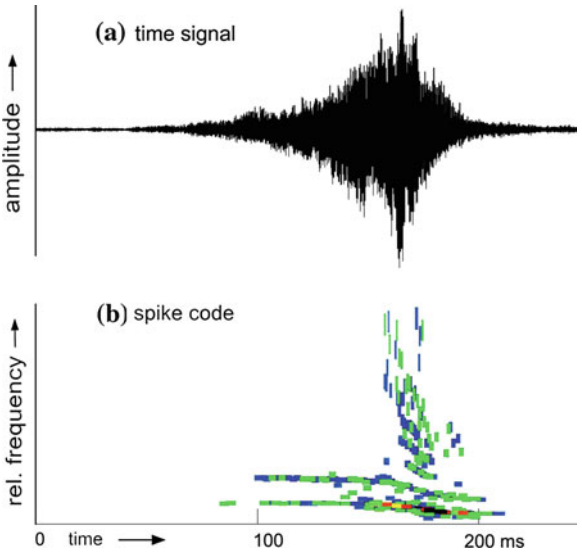


Fig. 3 **a** Time-domain signal of a *whoosh* sound, **b** corresponding event-based representation using Gammatone filters. Each *rectangle* corresponds to one basis function in the expansion of (1). The magnitude of the coefficient, a_k, is represented by the *gray shade* (darker → larger). Time, t_k, and center frequency, f_k, are represented by the location of the *rectangles*. The size of the *rectangles* indicates the localization of the basis functions in time-frequency space

that has been labeled *"whoosh"*. A Gammatone filter bank of 256 filters is used for generating a representation based on $K = 32$ atomic events. The salient areas in the signal are represented by more events than other areas.

Different auditory events can now be compared via their event-based representations by using appropriate distance or similarity measures, respectively. These distance measures can account for invariances in a straightforward way and can be devised such that they capture the event-structure of a complex auditory event [2]. Take, for example, the representation shown in Fig. 3b. The total distance between two such representations can, for example, be decomposed into a sum of distances between pairs of corresponding atomic events which, in turn, can be calculated as a weighted sum of the distances between the parameters describing them. But how can corresponding events be found? They can be found by minimizing the total distance over all the candidate pairs. The problem of optimally matching two event-based representations can be transformed into the problem of optimally matching the two node-sets of a weighted bipartite graph, which can in turn be solved using the so-called *Hungarian algorithm* [39]—details can be found in [2].

Supervised Learning: Support-Vector Classification

Support-vector learning is an efficient machine-learning method for learning the parameters of perceptrons for classifying patterns but also for assigning real-valued attributes, for example the quality ratings for auditory events event. Consider a simple binary classification problem, where data points represented by feature vectors, $\mathbf{x}$, —for example, some low-level descriptors of sound events-should be assigned to one of two possible classes, $y \in \{-1, +1\}$. Support-vector learning is a supervised learning method, hence it requires a so-called training set of labeled examples, that is, of pairs, $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, p$, during the learning phase. The *support-vector classifier* is a standard perceptron and has the following simple form,

$$y(\mathbf{x}, \mathbf{w}) = \text{sign}\left\{ w_0 + \sum_{k=1}^{M} w_i K(\mathbf{x}_{i(k)}, \mathbf{x}) \right\}. \tag{3}$$

$\mathbf{w}$ is a vector of model parameters, and the data points, $\mathbf{x}_{i(k)}$, $k = 1, \ldots, M$, are data points from the training set.

During the learning phase, model parameters have to be determined and specific data points, $\mathbf{x}_{i(k)}$,—the so-called *support vectors*—have to be chosen from the training set, such that the resulting classifier will perform well during recognition. Details about standard support-vector learning can be found in a number of textbooks, for instance [46]. Important for the following, however, is the function $K(\mathbf{x}_i, \mathbf{x}_j)$, the *kernel*, which is part of the perceptron. It can be interpreted as a similarity measure that quantifies how similar two data points, $\mathbf{x}_i$ and $\mathbf{x}_j$, are. Learning and recognition make both use of similarity values $K(\mathbf{x}_i, \mathbf{x}_j)$ for pairs of data points only, and feature values would only enter these processes through the similarity measure K.

Therefore, the function $K(\mathbf{x}_i, \mathbf{x}_j)$, which maps pairs of feature vectors to similarity values, can be replaced by a more general function, $K(i, j)$, that maps pairs (i, j) of patterns directly to similarity values. This allows support-vector learning and perceptron recognition to directly operate on relational rather than on feature-based representations.

Now consider event-based representations and the example of Fig. 3 . For every pair of event-coded sounds, (i, j), the Hungarian algorithm can be used to determine their distance, $d(i, j)$, which can then be transformed into a similarity measure, $K_{i,j}$, for example, by use of the common Gaussian kernel function

$$K(i, j) = \exp\left(-\frac{d(i, j)^2}{2\sigma^2}\right). \tag{4}$$

σ^2 is a variance parameter to be determined—sometimes called length scale. Unfortunately, this choice may not lead to a valid kernel function, valid in the sense of standard support-vector learning, as the kernel function should be positive semi-definite. Consequently, variants of support-vector learning have to be used that do not require this property [34, 35].

In [2] event-based representations and support-vector classification have been applied to recognize everyday sounds. The dataset contained ten classes of different types of everyday sounds. Figure 4 shows the results for a one-vs-the-rest recognition task, and compares the performance of the event-coded representation with the performance achieved for a number of standard feature-based representations for the same sound. The event-based representation, SPKE, outperforms the other representation schemes, including the popular feature-based representation using

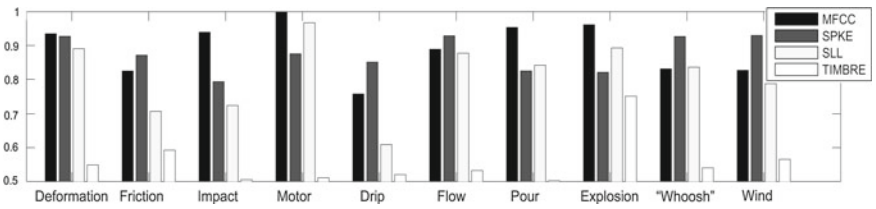


Fig. 4 Recognition of everyday sounds with perceptrons and support-vector learning. The figure shows the recognition performance for one particular sound class against the rest for a sound-data set consisting of ten sound classes. Four different low-level representations were used, mel-frequency cepstral coefficients, MFCC, an event-based representation using Gammatone filters, SPKE, a set of spectral low-level descriptors, SLL (energy, zero crossing rate, spectral centroid, roll-off frequency and their variances, finite differences, and the variances of the differences) and a set of timbre descriptors, TIMBRE (perceptual spectral centroid, relative specific loudness, sharpness, roughness, signal autocorrelation, zero crossing rate, time frame, log attack time, temporal increase, decrease and centroid, effective duration, energy-modulation frequency, energy-modulation amplitude)

mel-frequency cepstral coefficients, MFCCs,[2] in five particular cases. In the other five cases, MFCCs perform best.

As a result, the event-based representation can be considered as a denoising procedure that emphasizes those contours of a given sound that are perceptually important. On the other hand, however, from a certain number of atoms selected, some perceptually non-existent contours (phantom-spikes) can be emphasised. This could degrade the recognition performance. By means of the results of some suitable psycho-acoustical experiments, the perceptually optimal number of spikes can be determined to avoid this effect.

Furthermore, the total distance computed between two spike codes is a weighted sum of the parameters. These weights can be adjusted using prior knowledge. For impact sounds, for example, time differences can be weighted more strongly than differences in amplitude or frequency, which in turn improves recognition performance.

Unsupervised Learning: Prototype-Based Clustering

The use of relational representations is not limited to supervised learning paradigms and support-vector learning but can be applied to unsupervised learning as well. Although in the framework as laid out in Fig. 1, unsupervised learning is generally better suited for learning feature representations, clustering could also be a promising method for defining proto-objects—provided that the quality of the pre-segmentation is sufficiently high. If applicable, unsupervised learning has the benefit of not requiring annotations of auditory objects during the learning phase, since these are often expensive to obtain.

During clustering, data points are grouped according to a predefined similarity or distance measure. A *cluster* is then formed by data points whose inter-point distances are small compared to the distances to data points that are members of the other clusters. For some of the methods, a prototypical data point is generated for every cluster. These methods are usually called central or prototype-based clustering methods. In the following, it will be illustrated how prototype-based clustering methods can be applied to auditory objects in an event-based representation.

Let $d(i, j)$ be the distance between two auditory objects in their event-based representation as computed, for example, by using the distance measure introduced in the preceding subsection. A particular clustering of a set of auditory events can then be quantified using binary assignment variables, M_{ib}, where

$$M_{ib} = \begin{cases} 1 & \text{if the auditory event, } i, \text{ is assigned to a cluster, } b, \text{ and} \\ 0 & \text{otherwise}. \end{cases} \tag{5}$$

[2] *Mel-frequency cepstral coefficients*, MFCCs, are the DCT coefficients of the logarithm of a mel-scaled signal spectrum. They have been introduced for the purpose of speech recognition [21], but have since proven versatile and found use in many other acoustic classification applications.

Let b denote the prototypical event-based representation of all the sounds assigned to group b. Then a good grouping and a good prototype should minimize the cost function,

$$\mathcal{H}^{oc}(\{b\}, \{M_{ib}\}) = \sum_{i=1}^{I} \sum_{b=1}^{B} M_{ib}\, d(i, b), \tag{6}$$

with respect to both the assignment variables, M_{ib}, and the parameters of the prototypical event-based representations, b. Following [29], the optimization can be performed via an expectation-maximization algorithm. This is an iterative procedure, where each iteration consists of two steps. In the first step, all distances, $d(i, b)$, are calculated and representations are preferably assigned to the group for which the distance to its prototype is smallest. In the second step, the parameters of the prototypical representations are chosen by minimizing (6), keeping the assignment variables fixed. For better convergence, a probabilistic version of this procedure is used in practice, where the binary assignment variables are replaced by assignment probabilities—details can be found in [29] or [1].

Figure 5 shows the prototypical event-based representations for three clusters from a sound class called *whoosh* on the left-hand side. The figure shows that the prototypes well represent the sounds assigned to a particular cluster and that information can be

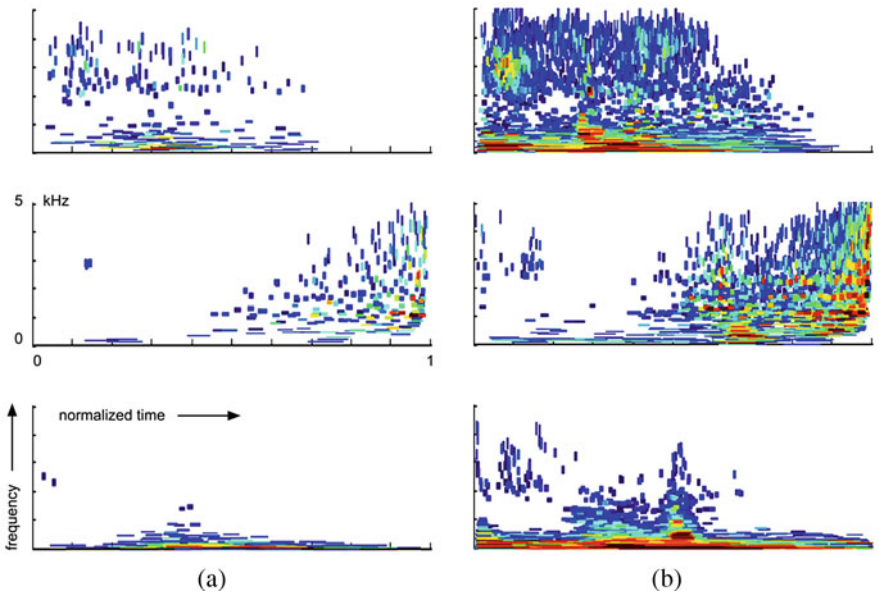


Fig. 5 **a** Event-based representations generated by prototype-based clustering of everyday sounds. Shown are three different prototypes for clusters of a class of sounds annotated by "whoosh". **b** Overlay of the event-based representations of all sounds as have been assigned to the corresponding prototype by the clustering method. For details of the graphical representation, see the caption of Fig. 3b

derived about the pattern of atomic events that are typical for a particular group. If annotations of sounds become available at this point, meaning can be assigned to the different groups, and further acoustic events can be recognized by matching them to the closest prototype.

2.2 Graphical Models as Dynamic Blackboards

At the highest level of the model framework depicted in Fig. 1, level (f), it is attempted to make sense of the input from level (e), which consists of sets of annotated proto-events and the confidence levels assigned to them. The task is to find out about the following.

- In how far does the input correspond to any patterns that are known to the system, that is, can any aural objects an/or aural scenes be *recognized*?
- How far can pre-known patterns be adjusted using available input information? In other words, can something be *learned* from the input—for instance, by taking advantage of any sort of understanding that we have of the interaction of aural objects?

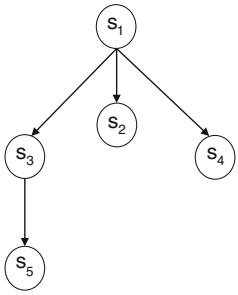
In this section, a mathematical tool is introduced that can be used to encode just such knowledge, namely, the so-called *graphical model*. Graphical models are originally models of a statistical nature, and they are typically designed on case-by-case information. They are thus generally ignorant of physical laws and mathematical/logical rules, such as acoustic-wave-propagation theory, the constancy of source identity, physical limitations to source and sensor movements, and other relevant knowledge.

To overcome these limitations of graphical models, it will now be attempted to reconcile the flexibility of graphical models with the precision of physical and mathematical knowledge—without loosing the advantages of either. To illustrate one way of obtaining an appropriate framework, the following three subsections will first present a brief background of graphical models and, consequently, "Auditory-Scene Understanding with Graphical Models" will show how rule-based knowledge can guide the design of graphical models to the end of achieving maximum precision with a task-specific limitation to relevant available knowledge.

Graphical Models

The dependence relationships within groups of random variables are often described very concisely by denoting their statistical dependence with the help of a graph. Such graphs come in two forms, directed and undirected. Yet, in the following, the discussion will be limited to the directed ones. In this type of graphs, two variables that are statistically dependent upon each other are connected with a line, with the arrow pointing from the independent to the dependent variable.

Fig. 6 Dependence tree



An example of such a dependence graph is given by Fig. 6. In this example, the graph indicates that the two variables, s_3 and s_5, exhibit some form of statistical dependence, while there is no direct dependence between s_3 and s_2. Based on this understanding, the graph can be used to simplify the joint probability density of the variables, s_1 through s_5, as follows,

$$\begin{aligned} \mathrm{p}(s_1, s_2, \ldots s_5) &= \mathrm{p}(s_1)\mathrm{p}(s_2|s_1)\ldots\mathrm{p}(s_5|s_1\ldots s_4) \\ &= \mathrm{p}(s_1)\mathrm{p}(s_2|s_1)\mathrm{p}(s_3|s_1)\mathrm{p}(s_4|s_1)\mathrm{p}(s_5|s_3). \end{aligned} \tag{7}$$

More generally speaking, a tree-shaped dependence graph indicates that the joint-probability-density function, PDF, of all variables, $\mathrm{p}(s_1, \ldots s_n)$, can be factorized in the following manner,

$$\mathrm{p}(s_1, s_2, \ldots s_N) = s_r \prod_{n\in N, n\neq r} p(s_n|s_{\mathrm{A}(n)}), \tag{8}$$

where r denotes the root node of the graph and $\mathrm{A}(n)$ yields the ancestor (or parent) nodes of node n. Graphs that are more complex can also be factored according to the same principle, that is, by using the fact that the probability density of statistically independent variables may be factorized into a product of the PDFs of all interdependent subgroups.

Graphical Models for Non-Stationary Processes

In many contexts, graphical models denote statical-dependence relationships, as is also true in the example shown in Fig. 6. It is, however, in the nature of auditory events that they are temporally evolving, and exhibiting only approximate short-time stationarity. Thus, when a description of auditory scenes is desired, it becomes necessary to extend graphical models to describe temporally evolving variables in addition to stationary ones.

One example of such a temporal graphical model is a *hidden Markov model*, HMM, which is highly popular due to its flexibility and the availability of easily

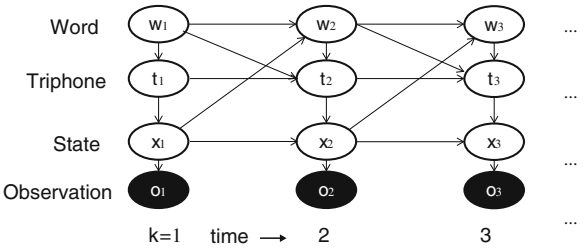


Fig. 7 Hidden Markov model, shown as an unrolled temporal graphical model

implementable, statistically optimal algorithms. Figure 7 depicts an HMM in the notation of a directed graphical model.

In this figure, variables are shown either as filled ellipses, meaning that they are observable, or as white ellipses, indicating that they are hidden. The exemplary variables here are taken from automatic speech recognition, ASR. They indicate the word identity, w_k, the triphone identity, t_k, the state index, x_k, and the observation vector, o_k, for each time frame, k.

As above, their dependence relationships are depicted in the notation of graphical models, but now, in addition, their temporal evolution is shown, because in contrast to the stationary variables in Fig. 6, it becomes necessary to model their value at each time frame by a separate node.

This step of adding duplicates of variables for each of the points in time is often described as *unrolling* the graphical model over time. As visible in this example, an originally static dependence graph can be extended by temporal unrolling such as to model the properties of a temporally-evolving, non-stationary statistical process.

Rule-Bases in Graphical Models

While graphical models, both for stationary and for non-stationary processes, are statistical models by nature, they allow for the incorporation of rule-based knowledge in two distinct ways, that is,

- On the one hand, conditional probabilities in the graph can also collapse to deterministic ones—namely, conditional-probability tables can contain ones and zeros—if physical rules or rules based on other knowledge sources state that certain values in one variable of the model directly imply setting other variables to specific values.
- On the other hand, the structure of the graphical model itself, or of sub-structures in it, can be compiled rather than determined manually, and this compilation process can be carried out by automatically integrating sets of rules. An example of how this may be accomplished in the context of automatic speech recognition by compiling possible sequences of phonemes from a pronunciation dictionary and a task grammar, can be found in [8].

Auditory-Scene Understanding with Graphical Models

By using graphical models like those described above, very precise source models may be attained, provided that training takes place on sufficient amounts of representative data. Such models can serve as a repertoire of possible source signals for schema-based, rather than primitive grouping. As one success story of graphical models in multi-source scene understanding, Hershey et al., [33], use factorial hidden Markov models as models and apply them to perform inference with variational loopy-belief propagation. The computational effort of this approach scales linearly in the number of sources and leads to super-human ASR rates on multi-speaker single-channel recordings.

This is already a highly promising result for the application of graphical models in aural-scene understanding. However, the current approach as exemplified by this work, still has several drawbacks, such as the following. The model of [33] and other similar graphical-model-based systems like [22] need to exploit very strong source models, that is, top-down or schema-based segregation, in order to come close to or even surpass the recognition rate of human listeners on the tested but highly specialized task. In natural scenes, there is not usually such an ample experience regarding the possible waveforms of sound sources. However, in any realistic cases, a much higher variability of possible sounds, coming with a far greater vocabulary and a much wider range of admissible sources, is the standard.

Thus, dynamical Bayesian networks have been applied and already proven their merit in cases where detailed source models were available. Yet, achieving a general applicability with this approach is still an open issue.

As a step forward, even for quite general tasks where only coarse source models are available, additional information in two specific forms could often be exploited.

- Physical and mathematical knowledge may be used, for instance, regarding the mixing process—which may be supported by input from other modalities.
- Psychoacoustic heuristics for source separation may be applied, as based on an implementation of *primitive-segregation* rules mimicking human auditory streaming in general environments.

How these two information sources, namely, physical models and perceptual rules, should be combined with all available source information, for the purpose of gaining an optimally informed understanding of auditory scenes, is an interesting open issue.

As one option, this should be possible by compiling a combination model from all information sources—similarly to well known model compilations for ASR, where a lexicon, phonetic and linguistic information are used to form a search network. Yet, for the applications envisaged here, the compiled model would not be linear in topology but rather allow for superposition of all acoustic sources according to the internal acoustical wave-propagation sub-model.

Such a compiled graphical model could also possess an interface for higher level processes that might search over variable allocations in the manner of an expert system. Only, in contrast to standard expert systems, this search would operate on the graphical-model variables directly, effectively making the graphical model an *active*

blackboard, which could be simulated to measure the goodness-of-fit between all observations and the internal variable occupation probabilities, corresponding to possible internal scene interpretations. Thus, of all scene interpretations, the most fitting one could be selected, whereby the "fit" is assessed, among other information, based on what is known about the source and further physical and mathematical knowledge, and guided by streaming mechanisms as are also active in human perception.

3 Active Learning and Active Listening

3.1 Active Learning

Active learning, that is, autonomous inductive learning, is one of the key features of the envisaged artificial listener and plays a prominent role in most of the higher-level processing stages of the framework that has been laid out in Fig. 1. Ideally, the artificial listening system would autonomously explore its environment and use the information gathered through this interaction in the learning process. Strategies for autonomous learning exist in principle, and reinforcement learning is a prominent example of this. However, *pure* autonomous learning is still notoriously slow in complex environments, and one has to resort to supervised learning strategies for many of the subtasks involved. Still, an artificial listener is an excellent testbed for concepts to improve *autonomy*.

Supervised learning suffers from the fact that acoustic events have to be annotated. Given the large amount of data needed by standard learning methods when tasks become complex, the required human interaction can become excessive. For an artificial cognitive system it is therefore important to make best use of the available information. One idea, which has been around for several decades by now, is to replace a passive *scanning* of the environment by strategies where a learning system *actively* sends out requests for training data that are particularly informative. There is a large amount of empirical evidence about the fact that *active data selection* is more efficient in terms of the required number of training examples for reaching a particular level of performance. It follows, that the amount of human interactions can be reduced by supervised-learning paradigms. The next subsection provides an example where active data selection is applied for learning a predictor for perceptual attributes assigned to everyday sounds by humans.

Active Data Selection

For a binary classification problem, consider a parameterized family of perceptron classifiers,

$$y(\mathbf{x}, \mathbf{w}) = \text{sign}\left\{ w_0 + \sum_{k=1}^{M} w_i K(\mathbf{x}_{i(k)}, \mathbf{x}) \right\}, \tag{9}$$

similar to the classifiers that have already been introduced in the context of support-vector learning in Sect. 2.1. Assume that the reference-data points, $\mathbf{x}_{i(k)}$, for the functions K have been chosen in a sensible way. Then the goal of inductive learning is to find values for the parameters, w_i, such that the perceptron predicts the class memberships sufficiently well.

In active data selection, inductive learning is interpreted as a process, where predictors from the set—for example, perceptron classifiers from the parameterized family of (9)—are discarded if their predictions are inconsistent with the training data. Every new data point from the training set splits the current set of classifiers in two sets that differ in their prediction of the class label. Assume that the set of classifiers is endowed with a useful metric, for example, a metric taking into account that two classifiers are similar if their predictions are so too. Then a space of classifiers can be constructed, and volumes and distances can be defined. With those concepts one can then assess how useful a new data point is for training: A new data point is useful, given that the space of classifiers predicting membership of one class has about the same *size* as the other ones with regard to volume or maximum diameter of the corresponding subspace. At any stage during learning, an active-learning agent, when implementing, for example, the perceptron classifier, will select a useful data point and will ask for its class. When the information arrives, the agent will no longer consider the subset of classifiers with more or less disagreeing predictions but rather continue the learning process with those classifiers that have shown to predict correctly [32].

If there is no noise in the problem, and if the two classes can be separated in principle by a perceptron, active learning leads to an exponential decrease in the size of the set of consistent classifiers with the number of training data. For size meaning volume, for example, this follows from the fact that every well chosen new data point cuts the size of the set by half. Given a distance measure between classifiers that is related to their difference in prediction performance, the exponential reduction in volume then carries over to an exponential reduction of the classification error with training-set size. Unfortunately, this assertion may no longer hold if classes cannot be separated without errors. The reduction of classification error may then become polynomial again. Still, empirical evidence is abundant, that shows that active data selection strategies lead to a significant improvement of learning over standard inductive learning strategies.

For illustration, the kernel perceptron (9), has been applied together with active data selection for training an agent to predict the perceptual quality of sounds. Four classes of impact sounds were generated by an acoustic model and played to ten human listeners whose task was to rate them as *glass, metal, plastic* or *wood*. Sound-rating pairs were then used to train ten multiclass predictors based on the kernel perceptron, one for each listener—either using standard methods or active data selection. During standard training procedures, sound-rating pairs were randomly chosen for every new training sound while, during active data selection, the agent requested the most informative data point under the volume-reduction criterion—see Fig. 8a. For this demonstration, sounds were represented by the parameters of the acoustic models that were used for their creation, but a representation based on aural features

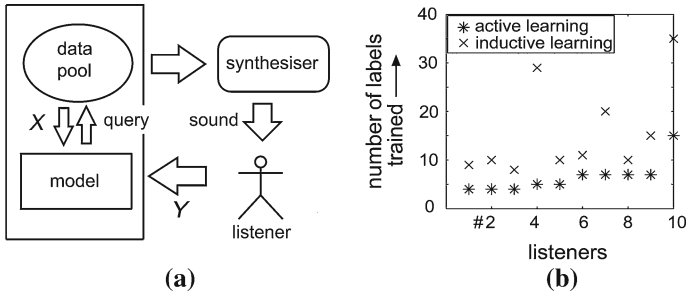


Fig. 8 Listening test using active learning. **a** A parameter vector, **x**, is used to synthesize a new sound by a physical model, that is, a synthesizer. After listening, the listener labels it with *glass, metal, plastic* or *wood* (label y). The learning machine then updates its prediction model and the query algorithm will suggests a new sound. **b** The graph shows the number of labeled training data required to reach a test-error rate of 0.35 for agents implemented through multiclass kernel-perceptron classifiers. The results for active data selection are compared with the results for standard inductive learning separately for every listener

could have been used as well. Figure 8b shows the number of training examples that were required to achieve a performance significantly above the chance level of 35 %. For all ten listeners, active learning led to a significant reduction of the number of training examples. Standard inductive learning required on average 2.5 times the number of labels compared to active learning.

3.2 Active Listening

The active-listening[3] approach, recently popular in robot audition [5], is based on the concept that perception and action come in couples [3]. While exploring their aural environment actively, listeners recognize sound sources and analyze aural scenes by simultaneously monitoring their auditory sensations and their motoric actions, taking advantage of both the auditory and proprioceptive modalities. Further, during this process, they permanently adjust their auditory system task-specifically. Modeling active listening implies various feedback mechanisms.

Feedback

Incorporation of multiple feedback loops into a model of sensory perception and cognition reaches out to the edge of current knowledge. In the auditory system, although there is strong evidence for numerous physiological feedback pathways

[3] The term *active listening* in the sense used here is not synonymous with a specific oral-communication technique that requires listeners to feed back to talkers what they hear.

[31, 45], little work has been done so far to incorporate feedback into processing models. In vision there is more experience with this type of modeling, but it is not yet clear to what extent this can be translated to the auditory domain [36, 49]. The aspect of porting domain experience across modalities has thus to be followed up carefully. While, in principle, the model framework of Fig. 1 allows for feedback between any stages, it makes sense at the time being to limit this structure with respect of physiological and operational evidence. In the following, a list of feedback paths is given that appear of particular relevance to the model framework presented here [15].

- Feedback from the binaural-mapping stage, that is, the output of auditory signal processing, to head-position control to keep a tracked sound source in aural focus—compare the so-called *turn-to reflex* as can already be observed in infants [19].
- Feedback from the cognitive stage to head-position in order to control for deliberate exploratory head movements, eventually moving the complete head-and-torso cart [16] this volume, and [7].
- Feedback from the segmentation stage to the signal-processing stage to solve ambiguities by activating additional preprocessing routines, for instance, cocktail-party effect and/or precedence effect processing—compare [50].
- Feedback from the cognitive stage to the signal-processing stage. This is intended to model efferent/reafferent effects of attention by modifying filter characteristics and/or putting a special focus on dominant spectral regions—for example [41, 42, 44, 47].
- Feedback from the cognitive stage to the segmentation stage to request task-specific and/or action-specific information on particular features—for example [18].

In the following, some exemplary feedback ideas along those lines are discussed in more detail.

1. To improve localization accuracy, a movable head-and-torso platform can perform movements, properly controlled by mimicking human strategies when exploring auditory scenes, for example, to derive estimates of distances and to solve front-back ambiguities [16], this volume.
2. Feedback can be used to adjust parameters for bottom-up processing, such as auditory-filter bandwidths, spectral weights in combining information across auditory filters, operating points of the temporal adaptation processes. Further, it can provide additional information supporting auditory-stream segregation, for instance, classifying groups of features of the same auditory stream in the binaural-activity maps [9].
3. At the *cognitive* level of the model framework, feedback from higher levels can make use of the interactive graphical-domain models as an *active blackboard*—as already mentioned in Sect. 2.2. Higher-level processes in an application-specific subsystem, such as an expert system for scene analysis, can set variables according to their specific intentions. Through modeling, it can be monitored how higher-level feedback corresponds with the rules and observations of the system and implications can be tested regarding the interpretation and initiation of new feedback information to lower model stages.

4. An important aspect of feedback is the incorporation of cross-modal information into the auditory processes. It is well known that profound interrelations between auditory and visual cues exist—compare, for example [38]. Visual cues can be introduced to the model system at two stages, namely, the pre-segmentation stage—the turn-to reflex—and the cognitive stage, exploiting prior knowledge about the visual scene. Specific proprioceptive information, such as the current position and movement of the head-and-torso platform, can also be used, particularly, at the pre-segmentation stage or even lower.

4 Conclusion

The model architecture as described in this chapter offers a comprehensive approach to modeling aural perception and experience. The listeners are modeled as an intelligent system exploring its surroundings actively and autonomously via an *active exploratory listening* process. It is assumed that, in the course of this process, the perceptual and cognitive world of the modeled listeners evolves and differentiates. This notion of the essence of listening stands in contrast to a widespread view that sound signals impinging on listeners' ears are processed by their auditory systems in a purely bottom-up manner. While it is an advantage of the latter approach that invention of the listeners is not needed, its prognostic power is limited to some primitive perceptual features, such as loudness, roughness or pitch. In a pure bottom-up approach, percepts are solely determined by the given ear-input signals, that is, formed in a signal-driven way. The active-explorative-listening approach, in contrast, requires a more complex model structure in which bottom-up, signal-driven, and top-down, hypothesis-driven, processes interleave in a complex way. The formation of hypotheses, a major feature of such a model structure, requires explicit knowledge inherent to the system. Part of the knowledge is acquired by the system itself in the exploration processes mentioned above, other knowledge has to be imported from external sources—potentially including physical knowledge as well statistical knowledge derived from possibly large datasets—or it may originate from other sensory modalities, such as proprioception, vision and/or tactility.

Acknowledgments The authors gratefully acknowledge suggestions of their external reviewers who helped to improve the clarity of presentation. Particular thanks are due to P. A. Cariani, who contributed relevantly by commenting the chapter from the viewpoint of biological cybernetics.

References

1. K. Adiloğlu, R. Annies, H. Purwins, and K. Obermayer. Deliverable 5.2, visualisation and measurement assisted design. Technical report, Neural Information Processing Group, TU Berlin, 2009.
2. K. Adiloğlu, R. Annies, E. Wahlen, H. Purwins, and K. Obermayer. A graphical representation and dissimilarity measure for basic everyday sound events. IEEE *Transactions Audio, Speech and Language Processing*, 20:1542–1552, 2012.

3. J. Aloimonos. *Active perception*. Lawrence Erlbaum, 1993.
4. M. Altinsoy. The quality of auditory-tactile virtual environments. *J. Audio Engr. Soc.,* 60:38–46, 2012.
5. S. Argentieri, A. Portello, M. Bernard, P. Danés, and B. Gas. Binaural systems in robotics. In J. Blauert, editor, *The technology of binaural listening*, chapter 9. Springer, Berlin-Heidelberg-New York NY, 2013.
6. L. Avant and H. Helson. Theories of perception. In B. Wolman, editor, *Hdb. of General Psychology*, pages 419–448. Prentice Hall, Englewood Cliffs, 1973.
7. M. Bernard, P. Pirim, A. de Cheveign, B. Gas, and IEEE/RSJ. Sensomotoric learning of sound localization from auditory evoked behavior. In: *Proc. Intl. Conf. Robotics and Automation*, ICRA ' 2012. pages 91–96, St. Paul MN, 2012.
8. J. Bilmes and C. Bartels. Graphical model architectures for speech recognition. *Signal Processing Magazine*, IEEE, 22:89–100, 2005.
9. J. Blauert. Analysis and synthesis of auditory scenes. In J. Blauert, editor, *Communication Acoustics*, chapter 1, pages 1–26. Springer, Berlin-Heidelberg-New York, 2005.
10. J. Blauert. Conceptual aspects regarding the qualification of spaces for aural performances. *Act. Acust./Acustica*, 99:1–13, 2013.
11. J. Blauert, ed. *The technology of binaural listening*. Springer, Berlin-Heidelberg-New York NY, 2013.
12. J. Blauert, J. Braasch, J. Buchholz, H. Colburn, U. Jekosch, A. Kohlrausch, J. Mourjopoulos, V. Pulkki, and A. Raake. Aural assessement by means of binaural algorithms - the AABBA project. In J. Buchholz, T. Dau, J. Dalsgaard, and T. Paulsen, editors, *Binaural Processing and Spatial Hearing*, pages 303–343. The Danavox Jubilee Foundation, Ballerup, Denmark, 2009.
13. J. Blauert and U. Jekosch. Concepts behind sound quality, some basic consideration. In *Proc. InterNoise 2003*, pages 72–76. Korean Acoust. Soc., 2003.
14. J. Blauert and U. Jekosch. A layer model of sound quality. *J. Audio-Engr. Soc.,* 60:4–12, 2012.
15. J. Blauert and K. Obermayer. Rückkopplungswege in Modellen der binauralen Signalverarbeitung (feedback paths in models of binaural signal processing). In *Fortschr. Akustik*, DAGA 2012, pages 2015–2016. Deutsche Ges.f. Akustik, DEGA, Berlin, 2012.
16. J. Braasch, S. Clapp, A. P. T. Pastore,, and N. Xiang. Binaural evaluation of auditory scenes using head movements. In J. Blauert, editor, *The technology of binaural listening*, chapter 8. Springer, Berlin-Heidelberg-New York NY, 2013.
17. A. Bregman. *Auditory scene analysis - the perceptual organization of sound*. MIT press, Cambridge MA, 1990.
18. N. Clark, G. Brown, T. Jürgens, and R. Meddis. A frequency-selective feedback model of auditory efferent suppression and its implication for the recognition of speech in noise. *J. Acoust. Soc. Am.*, 132:1535–1541, 2012.
19. R. Clifton, B. Morongiello, J. Kulig, and J. Dowde. Newborn's orientation towards sounds: Possible implication for cortical development. *Child develop.*, 52:883–838, 1981.
20. D. Corkhill. *Collaborating software: blackboard and multi-agent systems and the future*. Proc. Intl. Lisp Conf., New York NY, 2003.
21. S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE *Trans. Acoust., Speech, Signal Processing*, 28:357–366, 1980.
22. M. Delcroix, K. Kinoshita, T. Nakatani, S. Araki, A. Ogawa, T. Hori, S. Watanabe, M. Fujimoto, T. Yoshioka, T. Oba, Y. Kubo, M. Souden, S.-J. Hahm, and A. Nakamura. Speech recognition in the presence of highly non-stationary noise based on spatial, spectral and temporal speech/noise modeling combined with dynamic variance adaptation. In *Intl. Worksh. Machine Listening in Multisource Environments*, CHiME 2011, pages 12–17, 2011.
23. L. Dempere-Marco, D. Melcher, and G. Deco. Effective visual working memory capacity: An emergent effect from the neural dynamics in an attractor network. PLoS ONE, 7:e42719, 2012.
24. S. Deneve. Bayesian spiking neurons I: Inference. *Neural Computation*, 20:91–117, 2008.
25. S. Deneve. Bayesian spiking neurons II: Learning. *Neural Computation*, 20:118–145, 2008.

26. DIN EN ISO 9000. *Qualitätsmanagementsystem, Grundlagen und Begriffe (quality management system, fundamentals and concepts).* Dtsch. Inst. f. Normung, Berlin, 2005.
27. R. Engelmore and A. Morgan (eds.). *Blackboard systems*. Addison-Wesley, Boston MA, 1988.
28. L. Erman. The Hearsay II speech-understanding system - integrating knowledge to resolve uncertainty. *Computing surveys*, 12:213–253, 1980.
29. S. Gold, A. Rangarajan, C.-P. Lu, and E. Mjolsness. New algorithms for 2d and 3d point matching: Pose estimation and correspondence. *Pattern Recognition*, 31:957–964, 1998.
30. S. Haykin. *Neural networks - a comprehensive foundation*. Macmillan, New York NY, 1994.
31. J. He and Y. Yu. Role of descending control in the auditory pathway. In A. Rees and A. Palmer, editors, *Oxford Hdb. of Auditory Science*, volume 2: The auditory brain. Oxford Univ. press, New York NY, 2009.
32. F.-F. Henrich and K. Obermayer. Active learning by spherical subdivision. *J. Machine Learning Res.*, 9:105–130, 2008.
33. J. R. Hershey, S. J. Rennie, P. A. Olsen, and T. T. Kristjansson. Super-human multi-talker speech recognition: A graphical modeling approach. *Comput. Speech Lang.*, 24:45–66, 2010.
34. S. Hochreiter, T. Knebel, and K. Obermayer. An SMO algorithm for the potential support vector machine. *Neural Computation*, 20:271–287, 2008.
35. S. Hochreiter and K. Obermayer. Support vector machines for dyadic data. *Neural Computation*, 18:1472–1510, 2006.
36. B. Julesz and I. Hirsh. Visual and auditory perception - an essay of comparison. In E. Davis jr and P. Denes, editors, *Human communication* - a unified view, pages 283–340. McGraw Hill, New York NY, 1972.
37. A. Kohlrausch, J. Braasch, D. Kolossa, and J. Blauert. An introduction to binaural processing. In J. Blauert, editor, *The technology of binaural listening*, chapter 1. Springer, Berlin-Heidelberg-NewYork NY, 2013.
38. A. Kohlrausch and S. van de Par. Audio-visual interaction in the context of multi-media applications. In J. Blauert, editor, *Communication Acoustics*, pages 109–134. Springer, Berlin-Heidelberg-New York NY, 2005.
39. H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97, 1955.
40. S. G. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. IEEE *Transactions Signal Processing*, 41:3397–3415, 1993.
41. R. Meddis, R. Ferry, and G. Brown. Speech innoise and the medial olovo-cochlear efferent system. *J. Acoust. Soc. Am.*, 123:3051–3051, 2008.
42. D. Messing, L. Delhorne, E. Bruckert, L. Braida, and O. Ghitza. A non-linear efferent-inspired model of the auditory system - matching human confusion in stationary noise. *Speech Communication*, 51:668–683, 2009.
43. R. D. Patterson and J. Holdsworth. A functional model of neural activity patterns and auditory images. *Advances in Speech*, Hearing and Language Processing, 3:547–563, 1996.
44. B. Scharf. Human hearing without efferent input to the cochlea. *J. Acoust. Soc. Am.*, 95:2813, 1994.
45. B. Schofield. Structural organization of the descending pathway. In A. Rees and A. Palmer, editors, *Oxford Hdb. of Auditory Science*, volume 2: The auditory brain. Oxford Univ. press, New York NY, 2009.
46. B. P. Schölkopf and A. J. S. AJ. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, 2002.
47. L. Schwabe and K. Obermayer. Learning top-down gain control of feature selectivity in a recurrent network of a visual cortical area. *Vision Research*, 45:3202–3209, 2005.
48. E. Smith and M. S. Lewicki. Efficient coding of time-relative structure using spikes. *Neural Computation*, 17:19–45, 2006.
49. R. Welch and D. Warren. Intersensory interaction. In K.R. Boff, L.Kaufmann, and J. Thomas, editors, *Hdb. of Perception and Human Performance*, chapter 25, pages 1–36. Kluwer Academic, Dordrecht, 1989.
50. S. Wolf. *Lokalisation von Schallquellen in geschlossenen Rumen (Localization of sound sources in enclosed spaces).* doct. diss., Ruhr-Univ. Bochum, Germany, 1991.