Feedback loops in engineering models of binaural listening

J Blauert, D Kolossa, Patrick Danès

Proceedings of Meetings on Acoustics

Volume 21 http://acousticalsociety.org/

167th Meeting of the Acoustical Society of America

Providence, Rhode Island
5–9 May 2014

Psychological and Physiological Acoustics: Paper 1pPP11

Feedback loops in engineering models of binaural listening

J. Blauert and D. Kolossa, Ruhr-Universität Bochum, Bochum, Germany, jens.blauert@rub.de, dorothea.kolossa@rub.de

P. Danès, LAAS–CNRS and Université de Toulouse III – Paul Sabatier, Toulouse, France, patrick.danes@laas.fr

Hearing models for tasks like auditory scene analysis or sound-quality judgments can run into severe problems when acting in a purely bottom-up, that is, signal-driven manner, as they may have to follow up on all possible output options until a final decision is taken. This may lead to a combinatorial explosion. A way out is the inclusion of top-down, that is, hypothesis-driven processes. In top-down processing, the number of states to be evaluated can be reduced substantially when the system *knows what to look for* and thereby focuses attention on states which *make sense* in a given specific situation. To implement adequate top-down processes, we include various feedback loops in our current hearing model, some specific, others more general. The general ones originate from the concept that the listener model (*"artificial listener"*) actively explores acoustic scenes and thereby develops its aural world in an autonomous way. Following this notion, listeners are modeled according to the *autonomous-agents* paradigm, where agents *actively learn and listen*. [Work supported by EU–FET grant TWO!EARS, ICT–618075, <www.twoears.eu>.]

1. Introduction

In psychology, at least since Gibson (1966), it is no longer considered adequate to conceive the auditory system as a receiver that simply projects the "external" world onto the central nervous system, so that a more or less imperfect internal representation of the external world can be formed. In fact, modern concepts start from the notion that the nervous system forms its world in an interactive explorative process, in the course of which the aural world develops and differentiates. As to the structure of the auditory system, this requires that, besides pure *bottom-up* (signal-driven) processes, *top-down* (hypothesis-driven) processes are assumed. Models of the binaural system that take this situation into consideration have been proposed before (e.g., Blauert 1999, Blauert *et al.* 2010). A further extension of these models is required due to the active exploratory character of the auditory processes. Namely, *feedback loops* have to be included, thus essentially making the models "*cybernetic*". The necessity for including feedback becomes particularly clear when considering robot audition, for example, a robot being assigned the task of autonomously exploring acoustic scenes (Raake *et al.* 2014).

It is a well-known fact that humans, attending to a sound, move their head into the assumed direction of the sound source in a reflexive (*turn-to reflex*) or reflective manner. Obviously, this requires feedback from the auditory system to the motor system (Bernard *et al.* 2012). It is further known that the highest physiological stage of the auditory system, that is, the auditory cortex, projects back to all lower levels of auditory processing (Schofield 2009, He & Yu 2009).

Our system (the TWO!EARS system) is being built on the basis of a movable robot with an artificial head. Its ear signals are preprocessed in a number of steps that mimic corresponding steps in the human auditory system. As a result, a binaural activity map is rendered as an intermediate representation. The binaural map is then evaluated by a segmentation and labeling process which is based on a variety of rule- and data-oriented algorithms. The following processing stages contain not only statistical knowledge, but also knowledge imported from external experts. In this way, context information is integrated and meaning can be assigned to the data, which are finally present in form of symbols. Depending on the specific application, scenes are analyzed, actions are triggered, and/or quality judgments are performed – just to name some exemplary application opportunities. The "intelligent" part of the model architecture is organized in a so-called "blackboard" architecture (Engelmore & Morgan 1988, Kolossa 2011) – see Sec. 2 for details.

2. Feedback loops under consideration

In the following, we list a number of assumed feedback loops, which are deemed relevant for technical applications of binaural models. They have been discussed in the context of the architecture of the TWO!EARS system (Raake *et al.* 2014, see also Blauert & Obermayer 2012, Blauert *et al.* 2013). Although the TWO!EARS architecture allows for feedback loops between all processing levels of the model structure, a selection for the actual implementation in the project has been made on the basis of functional relevance. The selected feedback loops are:

- Feedback from the binaural mapping stage, namely, from the auditory signal processing stage, to control the position of the head (e.g., the *turn-to reflex*).
- Feedback from the cognitive stage to head-position control for exploratory head movements. To improve localization accuracy and to solve front-back ambiguities, the model head performs movements, properly controlled by mimicking human strategies when exploring aural scenes. Further, obstacles that cause acoustic occlusion can thus be recognized and circumvented.

- Feedback from the segmentation stage down to the signal-processing stage to solve ambiguities by activating additional specialized preprocessing routines, such as *cocktail-party effect, precedence effect or de-reverberation* processing.
- Feedback to change processing parameters in the bottom-up stages, like changing spectral weights in combining information across auditory filters, adjusting the operating point of the temporal adaptation processes (*olivo-cochlear effect*), or providing additional information that supports auditory-stream segregation – for example, by classifying groups of features within the activity maps as belonging to the same auditory stream (*Gestalt* rules).
- Feedback from the cognitive stage to the segmentation stage, such as requesting task-specific or action-specific information on particular features – further to suppress information less relevant for the specific task (*information masking*).

Two particularly relevant feedback structures are exemplarily discussed in the following.

3. Feedback loops within the blackboard

At the "cognitive" level of our model system, feedback from higher levels is integrated by using a graphical model as active blackboard architecture (Fig. 1). Higher level processes in application-specific subsystems, such as a software expert of scene analysis, can set variables according to their particular intentions. Then, after an inference in the graphical model has been carried out accordingly, it becomes visible how higher-level feedback corresponds with the rules and observations of the system, and what implications can be drawn from it. The graphical models stem from multiple sources of information and are composed on the blackboard to form one comprehensive description of the acoustic or audiovisual scene. The model parameters are adjusted in order to create a world model, namely, a description of the state of the environment which optimally matches all observations, that is, all sensor data that the cognitive system makes available. This structure allows for many types of feedback that can be initiated whenever the output of the system is not sufficiently reliable. Insufficient reliability is detectable within the graphical models, but care must be taken to distinguish continuous-valued variables like locations or intensities from discrete-valued variables like spoken words or source identities.

- If there is insufficient reliability in a continuous-valued variable, this can be seen from large variances of the estimate. For instance, if the system is tracking an acoustic source, high variances of the location estimate are indicative of an unreliable interpretation, as has been successfully exploited in Schymura *et al.* (2014).
- For discrete-valued variables, confusions are detected when there are multiple interpretations that are assigned high likelihoods. One example where such problems can occur, is given by situations with conflicting evidence, that is, when two or more contradictory interpretations are assigned high likelihoods by different subsystems. For example, one source maybe interpreted as a speaker by the acoustic model and as a radio by the visual model.

In both types of confusion, continuous-valued and discrete-valued, the graphical model architecture of the blackboard is helpful for triggering feedback and disambiguation. More specifically, we aim at using the connectivity of the complete graphical model for this purpose: When a variable on the blackboard is shown to have a high degree of uncertainty, the underlying causes of uncertainty are traced by following the dependency relationships of the variable backwards. This typically reveals one of the following three situations.

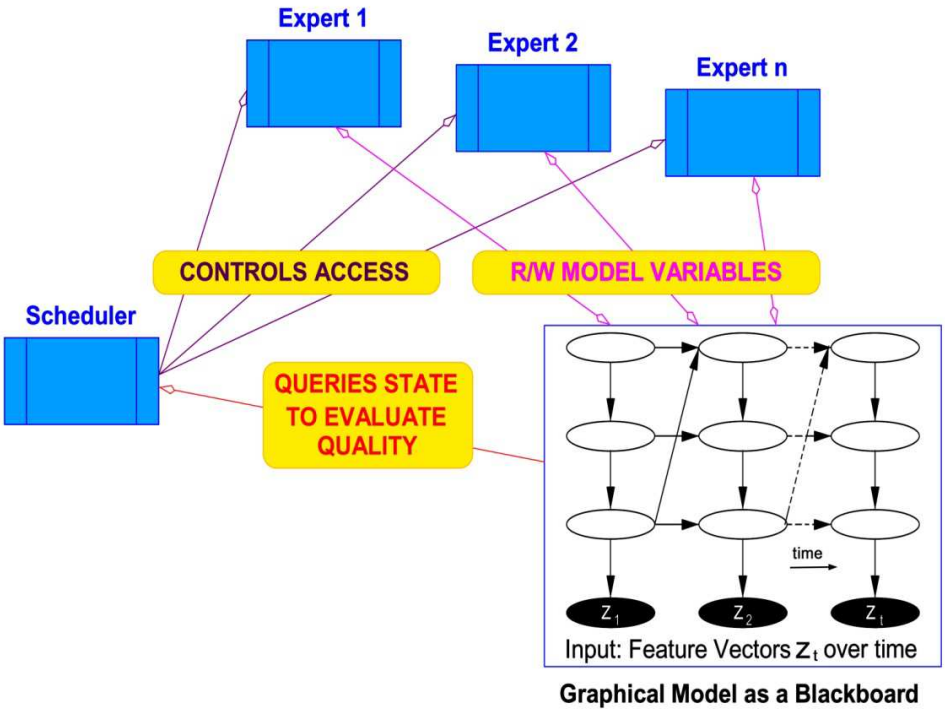

Fig. 1: Schematic plot of the proposed blackboard architecture

- *Ambiguities* Two or more causative variables make two different interpretations highly likely. A typical reason for this issue could be found in conflicting inputs from multiple modalities. One method for resolving such conflicts is an appropriate fusion of modalities, guided by variances or by reliability indicators like observation uncertainties (Vorwerk 2011) of all modalities. These reliabilities or uncertainties are partially available from prior knowledge but should be adjusted to the situation as far as possible, for example by model adaptation.
- *Surprise* Prior knowledge suggests an interpretation that is different from the interpretation suggested by current sensory input. If prior and evidence are in conflict, that is, if expectations and observations are inconsistent, two options exist, namely, either adjust the prior or adjust the interpretation. Making the decision of which of the two to favor will be driven by Bayesian information criteria such as Minimum Description Length (MDL).
- *Uncertainties* One or more causative variables introduce high variances in the interpretation. In these cases, we follow the graphical-model (GM) connections backwards to identify the causes as far as possible. Depending on further available graphical models that can be included on the blackboard, we then decide whether to require additional inputs, such as additional features, additional views of the scene, or additional processing, for instance, adaptive de-noising, given that these promise an improvement of the accuracy of those variables that have been identified as the root causes of uncertainty.

4. Sensori-motor feedback loops

These feedback loops model hardwired behaviors that seamlessly interweave sensory stimulation and motion. They take place at the sensorimotor-reflex level on short time scales. A typical example in binaural audition is the aforementioned *turn-to reflex.* To a larger extent, the tight integration of motion and sensory stimulation complies with recent developments in embodied cognition (O'Regan & Noe 2001), postulating that our sensory experience arises from mastering the sensorimotor contingencies, that is, of how stimuli vary as a function of bodily movement. In robotics, the synthesis of so-called "active" binaural auditory functions, which incorporate the motor commands of the sensor, has long been acknowledged (Nakadai *et al.* 2000). The aim is to overcome limitations of their passive counterparts, such as front-back ambiguity and distance non-observability, or to perceive a source in the "auditory fovea" (Nakadai *et al.* 2002) while keeping the engineering design simple.

In the vein of Portello *et al.* (2014a), three fundamental stages have been identified, the first two being related to the analysis of the sensorimotor flow, the third being feedback in itself. These three stages comply with the blackboard architecture proposed in Sec. 2, and are defined as follows.

(1) *Short-term detection* Detection of the active sources and estimation of their spatial arrangement from the analysis of the binaural stream over small time snippets. The result is an input to the graphical model.

(2) *Audio-motor binaural localization* Assimilation of these data over time, and fusion with the motor commands of the sensor, so as to get a first level of active localization. A node of the graphical model is used to store the posterior pdf, $p(x_k|z_1...z_k)$, of the corresponding location, x_k, given all observations, $z_1...z_k$.

(3) *Information-based feedback control of the binaural sensor* Feedback control, which delivers adequate sensor motor commands in order to improve the fusion performed in active localization, carried out by an expert in the architecture.

The extraction of spatial source characteristics in Stage (1) is, for example, performed through maximum likelihood estimation under the assumptions that the sound-source and noise signals are jointly Gaussian locally stationary random processes (Portello *et al.* 2014b). This approach fully takes into account scattering effects. It assumes negligible relative motion between the binaural sensor and the sources, as well as prior knowledge of the noise statistics. While a closed-form separable solution can be obtained for the single-source case, multiple sources are handled via the *expectation-maximization algorithm* if they are *W-disjoint orthogonal*, that is, if one source at most has significant energy in each "bin" of the channel-frame-frequency decomposition of the binaural stream. Source activity detection is addressed through information criteria.

The assimilation of the history of outputs from Stage (1) over time, and its fusion with the motor commands of the sensors, is performed naturally in a stochastic filtering scheme, which is the cornerstone of Stage (2) (Portello *et al.* 2014a). The state vector of the underlying stochastic state-space model, that is, the internal vector variable $\boldsymbol{x}$, which is the subject of estimation, describes the sensor-to-source situation. The control input to this model is constituted by the motor commands of the binaural sensor.

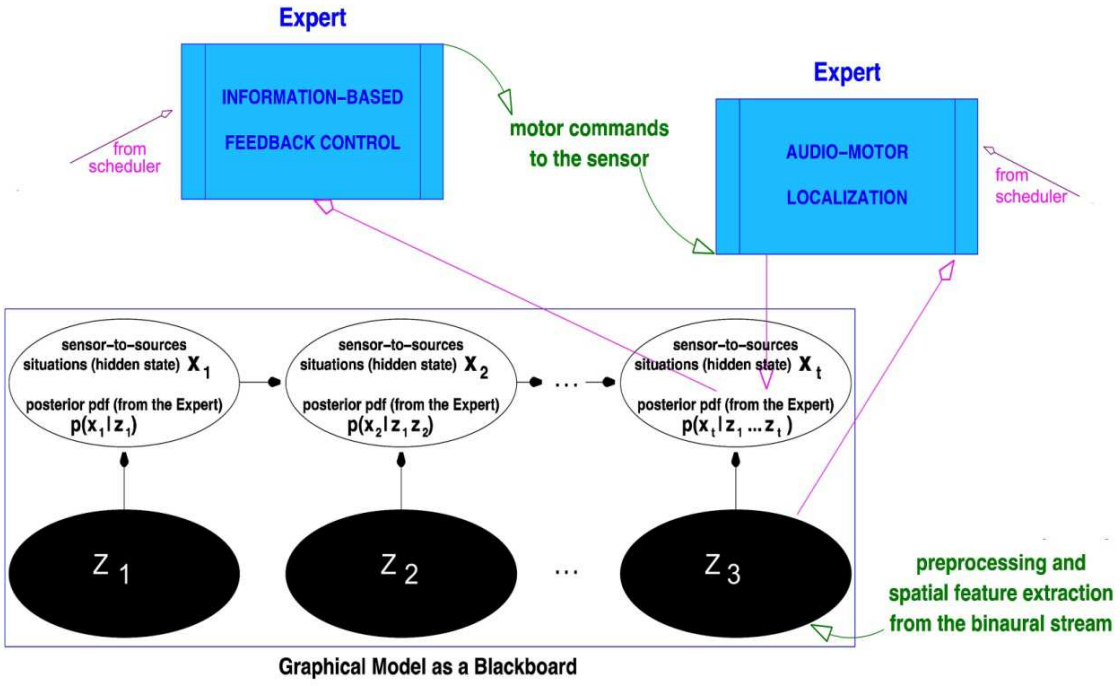

Fig. 2: Schematic plot of feedback at the sensorimotor level

The stochastic state equation then depicts the effect of sensor motion on localization, for instance, through rigid body kinematics. If the sources also move, then their absolute positions must be inserted in the unknown state vector, and their dynamics are described by an autonomous system with unknown initial condition.

The challenges are threefold: *First*, due to model nonlinearities, the consistency of the filter must be carefully examined. Even if the filter relies on perfectly known noise statistics, the approximation of the state posterior density function (pdf) that it delivers can be overly optimistic or inconsistent due to overestimation of the range, etc. *Second*, the filter must be endowed with self-initialization as well as with routines for handling false measurements and source intermittence. Data association problems predictably occur in the case of multiple sources. *Third*, the motion of the sensor obviously affects the quality of localization. As mentioned, Stage (3) is addressed through the design of the feedback controller (*Fig. 2*). The idea is to define a criterion that judges the quality of exploration based on the parameters of the posterior pdf of the state. If the synthesis of the control law is guided by this factor, then other competing objectives have to be included, as, for instance, the energy of the control signal. The challenge is to bridge the gap between the mathematical statement of the problem and a tractable implementation.

4. Conclusions

We have presented a framework for modeling active binaural listening and discussed avenues for implementing feedback within this architecture. The feedback is guided by the principle that the system should be able to focus on estimating task-dependent quantities of interest with the maximum possible accuracy. We have described approaches for optimizing towards this goal, taking into consideration all stages of our system, from signal preprocessing up to the cognitive

stages. The integration of all sources of prior and current information is driven by a graphical-model-based world model, which considers all estimates and information sources as random variables. This allows us to weight all information according to its specific reliability, to trace dependency relationships, helping to identify the causative variables of any uncertainty in the system, and to carry out feedback control such as to minimize the variance of the variables that need to be estimated for the specific system task at hand.

References

Bernard, M., Pirim, P, de Cheveigné, A. & Gas, B. (2012). Sensomotoric learning of sound localization from auditory evoked behavior. *Proc. IROS 2012*, P–Vilamoura

Blauert, J. (1999). Models of binaural hearing: architectural considerations. *Proc. 18th DANAVOX Symp.*, 189–206, Danavox Jubilee Found., DK−Ballerup

Blauert, J., Braasch, J., Buchholz, J., Colburn, H.S., Jekosch, U., Kohlrausch, A., Mourjopoulos, J., Pulkki, V. & Raake, A. (2010). Aural assessment by means of binaural algorithms – the AABBA project. In: Buchholz, J.M., Dau, T., Dalsgaard, J.C. & Poulsen, T. (eds.) *Binaural Processing and Spatial Hearing, Proc. 2nd Int. Symp. Auditory & Audiolog. Res., ISAAR'09*, 113–124, The DANAVOX Jubilee Found., DK−Ballerup

Blauert. J. & Obermayer, K. (2012) Rückkopplungswege in Modellen der binauralen Signalverarbeitung (Feedback paths in models of binaural signal processing), *Fortschr. Akust. DAGA'12*, 2015–2016, Dtsch. Ges. Akustik, D–Berlin

Blauert, J., Kolossa, D. Obermayer, K. & Adiloğlu, K (2013) Further challenges and the road ahead. In J. Blauert (ed.), *The Technology of Binaural Listening*, Springer, Berlin–Heidelberg–New York NY & ASA Press, New York NY

Engelmore, R.S. & Morgan, A. (eds.) (1988) *Blackboard systems*. Addison–Wesley, Boston MA

Gibson, J.J. (1966) *The Senses Considered as Perceptual Systems*. Houghton Mifflin, Boston MA

He, J. & Yu, Y. (2009). Role of the descending control in the auditory pathway. In: Rees, A. & Palmer A. R. (eds.) *Oxford Handb. of Auditory Science*, Vol. 2: The Auditory Brain. Oxford Univ. Press, New York NY

Kolossa, D. (2011). High-level processing of binaural features. *Proc. FORUM ACUSTICUM 2011*, DK-Aalborg

Nakadai, K., Lourens, T., Okuno, H.G., Kitano, H. (2000). Active audition for humanoids. *Proc. AAAI–2000*, Austin, TX

Nakadai, K., Okuno, H.G., Kitano, H. (2002). Exploiting auditory fovea in humanoid–human interaction. *Proc. AAAI–2002,* CN–Edmonton

O'Regan, J.K., & Noe, A. (2001). A sensorimotor account of vision and visual consciousness. Behavioral and Brain Sciences, 24(5), 939–1031

Portello, A., Bustamante, G., Danès, P., Piat, J. & Manhès, J. (2014a). Active localization of an intermittent source from a moving binaural sensor. *Proc. FORUM ACUSTICUM 2014*, PL–Kraków

Portello, A., Bustamante, G., Danès, P. & Mifsud, A. (2014b). Localization of multiple sources from a binaural head in a known noisy environment. *Proc. IROS'2014,* Chicago, IL

Raake, A., Blauert, J., Braasch, J., Brown, G., Danès, P., Dau, T., Gas, B., Argentieri, S., Kohlrausch, A. Kolossa, D., Le Goff,.N., May, T., Obermayer, K. and Spors, S. (2014). TWO!EARS – Integral interactive model of auditory perception and experience. *Fortschr. Akust. DAGA 2014, Dtsch. Ges. Akustik*. D–Berlin

Schofield, B. R. (2009). Structural organization of the descending auditory pathway. In: Rees, A. & Palmer A. R. (eds.) *Oxford Handb. of Auditory Science*, Vol. 2: The Auditory Brain. Oxford Univ. Press, New York NY

Schymura, C., Walther, T., Kolossa, D., Ma, N. & Brown, G. (2014). Binaural sound source localisation using a Bayesian-network-based blackboard system and hypothesis-driven feedback. *Proc. FORUM ACUSTICUM 2014*, PL–Kraków

Vorwerk, A., Zeiler, S., Kolossa, D., Fernandez Astudillo R. & Lerch, D. (2011) Use of Missing and Unreliable Data for Audiovisual Speech Recognition. In: Kolossa, D. & Haeb-Umbach, R. (eds.) *Robust Speech Recognition of Uncertain or Missing Data – Theory and Applications.* Springer, Berlin–Heidelberg–New York