7 - 12 SEPTEMBER 2014, KRAKÓW
FORUM ACUSTICUM

Binaural Sound Source Localisation using a Bayesian-network-based Blackboard System and Hypothesis-driven Feedback

Christopher Schymura, Thomas Walther, Dorothea Kolossa
Institute of Communication Acoustics, Department of Electrical Engineering and Information Technology, Ruhr-Universität Bochum, Germany

Ning Ma, Guy J. Brown
Speech and Hearing Research Group, Department of Computer Science, University of Sheffield, United Kingdom

Summary
An essential aspect of Auditory Scene Analysis is the localisation of sound sources in relation to the position of the listener in the surrounding environment. The human auditory system is capable of precisely locating and separating different sound sources, even in noisy and reverberant environments, whereas mimicking this ability by computational means is still a challenging task. In this work, we investigate a Bayesian-network-based approach in the context of binaural sound source localisation. We extend existing solutions towards a Bayesian network based blackboard system that includes expert knowledge inspired by insights into the human auditory system. In order to improve estimation of source positions and reduce uncertainty caused by front-back ambiguities, hypothesis-driven feedback is used. This is accomplished by triggering head movements based on inference results provided by the Bayesian network. We evaluate the performance of our approach in comparison to existing solutions in a sound-source localisation task within a virtual acoustic environment.

PACS no. 43.60.Jn, 43.66.Qp

1. Introduction

Human listeners have a remarkable ability to make sense of complex acoustic scenes, a phenomenon that has been termed *auditory scene analysis* (ASA) by Bregman [1]. Spatial hearing makes a substantial contribution to ASA, by allowing individual sound sources to be localised and perceptually segregated from other sounds (see [1, 2] for a review). Reproducing this ability in machine hearing systems is proving to be very challenging (for example, see [3]). In particular, current machine hearing systems are unable to localise sounds under conditions of noise and reverberation that present little difficulty for a human listener.

Machine hearing systems differ from human listeners in a number of important respects. The current paper focuses on two of these. First, machine hearing systems are typically implemented on a *static* platform, so that the acoustic sensors are in a fixed orientation. In contrast, human hearing is *active*; head movements provide listeners with information about changes in interaural time differences (ITDs) and interaural level differences (ILDs) which can be used to disambiguate the location of a sound source [4]. Secondly, machine hearing systems typically assume that information flow is strictly *bottom-up*. Again, this stands in contrast to auditory processing, in which *top-down feedback* is known to play an important role; in fact, there is evidence for pronounced top-down pathways in the human auditory system and also in the visual cortex [5, 6, 7]. Learning from the biological paradigm, it becomes clear that mere bottom-up feature processing cannot explain human capabilities in audiovisual analysis.

The current paper proposes a software architecture for machine hearing in which head movements and top-down feedback play a crucial role, which is being developed within the EU project TWO!EARS. Our approach is based on a *blackboard* problem-solving architecture, which was originally introduced in the HEARSAY-II Speech-Understanding System [8]. A blackboard system consists of a group of independent experts, also referred to as *knowledge sources* (KSs) that communicate by reading and writing data

on a globally-accessible data structure, the *blackboard*. The blackboard is typically divided into layers, corresponding to data, hypotheses and partial solutions at different levels of abstraction. Given the contents of the blackboard, each knowledge source indicates the actions that it would like to perform; these actions are then coordinated by a scheduler, which determines the order in which actions will be carried out. The blackboard architecture has a number of characteristics that make it eminently suitable for machine hearing: it provides a framework for reasoning about acoustic scenes that is flexible, opportunistic and integrates bottom-up processing with top-down feedback.

In the 1990s, a number of authors described blackboard-based systems for machine hearing [10, 11, 12, 13]. All of these systems were in most respects 'conventional' blackboard architectures, in which the knowledge sources consisted of rule-based heuristics. In contrast, the approach proposed here aims to exploit recent developments in machine learning, by combining the flexibility of a blackboard architecture with powerful learning algorithms afforded by probabilistic graphical models.

The remainder of the paper is organised as follows. Section 2 describes the bottom-up processing component of the Two!Ears architecture, which computes ITD and ILD cues from models of auditory processing. The graphical-model-based blackboard architecture is described in Section 3, where the motivation for it is also discussed in detail. Section 4 describes a methodology for evaluating the system on a single-source localisation task and presents the results. The paper concludes with general discussion in Section 5.

2. Bottom-up processing

2.1. Binaural signal generation

The binaural signals that serve as inputs to the auditory front-end are generated using head related transfer functions (HRTFs) obtained from a Kemar dummy head [14]. The HRTFs were recorded in an anechoic chamber with an angular resolution of 1° in the horizontal plane at a distance of 3 m from the source to the receiver. We use linear interpolation to obtain HRTFs corresponding to arbitrary angular positions. The left and right ear signals are then generated by filtering a single-channel source signal with the HRTF pair corresponding to the desired source location. Head movements are simulated by computing the relative angle between the target source position and the head orientation and adapting the HRTF interpolation to this specific angle.

2.2. Auditory front-end

To model the specific properties of the human auditory periphery, we use an auditory front-end that is adopted from [15]. The ear signals are processed by a bank of gammatone filters followed by inner-hair-cell processing. In order to model the frequency selctivity of the human basilar membrane, the gammatone filterbank consists of N fourth-order, phase-compensated gammatone filters. The center-frequencies of the filters are equally spaced on the equivalent rectangular bandwidth (ERB) scale [16]. Additionally, each gammatone filterbank channel is scaled with a specific gain to model the frequency response of the middle ear canal [17]. The gammatone filterbank output is further processed by applying half-wave rectification and square-root compression to account for the behavior of the inner hair cells. In this work, we apply a frame-based processing, dividing the incoming ear signals into overlapping frames, with a specific frame shift. The resulting signals serve as inputs to the blackboard system. Detailed parameters of the auditory front-end components used during the evaluation will be described in Section 4.2.

3. Blackboard architecture

The blackboard system proposed in this work is broadly based on the Hearsay-II Speech-Understanding System [8]. The central element of a blackboard system is the *blackboard* itself: it can be best described as a global data structure that represents knowledge that can be used to incrementally accomplish a certain task. Data that is stored on the blackboard can be manipulated by a set of *knowledge sources*. KSs collaborate via the blackboard by triggering when relevant data is available and depositing new data on the blackboard, which leads to a solution to the problem that should be solved. The architecture is event-driven; a change in the state of the blackboard (such as the arrival of new data) causes an event to be broadcast. A *blackboard monitor* is responsible for monitoring and handling these events. It maintains an *event register* that indicates which KSs should respond to a certain event. The possible actions that can be performed, given the current state of the blackboard, are listed in an *agenda*. A *scheduler* is then responsible for ranking possible actions and selecting one to perform. Completion of an action will most likely result in further changes in the state of the blackboard leading to broadcast of new events.

The design of the blackboard system allows for a fusion of statistical and expert knowledge. The novel approach we investigate in this work is the representation of knowledge by designing the blackboard as a set of interconnected graphical models, yielding a representation of the blackboard itself as a Bayesian network [9]. Computationally, this is realized by designing the blackboard to be a space for creating, assembling, and evaluating graphical models.

3.1. Motivation for a Graphical-model-based architecture

Graphical models have recently attracted great interest within the fields of machine learning and cognitive systems. They describe relationships between statistical variables in the form of simple graph structures. In these graphs, each node corresponds to a variable, and each edge indicates a dependency relationship between variables. In this way, graphical models can be used to describe the dependencies between all variables that are of interest, effectively providing a world model, which is not only mathematically useful but also interpretable.

Graphical models come in many different specific forms, such as Hidden Markov Models, Markov Random Fields, or dynamic state space models, which are suitable for creating precise descriptions of the constituent components of acoustic or audiovisual scenes. Efficient algorithms have been developed, which allow the optimal fit to be found between the model parameters and the observations taken from all sensors of a system. In effect, this means that, based on a graphical model of the audiovisual objects in an environment, the system will be able to find the best explanation of all available information, optimally fusing prior knowledge (e.g., linguistic or acoustic knowledge) with the currently available sensor input.

Taking graphical models as building blocks further allows us to

- consecutively build models of the audiovisual environment from smaller, well-understood models of environmental objects (including state-of-the art statistical models of auditory objects),
- understand sensory data as a composition of these source models and a model of the system's own "perception"
- and to understand the system's interpretation of the audiovisual environment, by virtue of the interpretability of each component and of their connections.

Since the model is statistical in nature, the resulting interpretation of the environment will not only denote the type, number, location and – if applicable – the possible intention of all objects of interest, but also contain estimates of the variances (or probability distributions) of all of these quantities. This will endow the system with the ability to judge the reliability of its own interpretation, and can ultimately be used to design active listening and active exploration, so as to ensure that the most relevant variables are determined with sufficient reliability.

3.2. Proposed blackboard architecture

Fig. 1 shows an overview of the general system architecture that is used in this work to solve a single-source localisation task. The blackboard workspace is arranged into a hierarchy of four layers:

The first and lowest layer, denoted as the *acoustic cues layer*, contains observation vectors modeled as continuous, multivariate and observable random variables. The observations are assembled of estimated ITDs and ILDs that can be added to the blackboard by the corresponding *Acoustic Cue KS* that operates on this layer. The Acoustic Cues KS takes the monaural left and right ear signals that were processed by the auditory front-end as inputs and estimates ITDs and ILDs independently for each frame and filterbank channel. The resulting observation vector

$$\boldsymbol{o}_k = \left(\hat{\tau}_{k,1},\, \ldots,\, \hat{\tau}_{k,N},\, \hat{\delta}_{k,1},\, \ldots,\, \hat{\delta}_{k,N}\right)^T \tag{1}$$

has $2N$ dimensions, where $\hat{\tau}_{k,l}$ denotes the estimated ITD at frame index $k \in \mathbb{N}_0$ and filterbank channel $l = 1, \ldots, N$ and $\hat{\delta}_{k,l}$ denotes the estimated ILD, respectively.

The central element of the second layer, which is referred to as the *location hypothesis layer*, is a discrete hidden random variable $\hat{\boldsymbol{\phi}}_k$ which represents hypotheses about the possible locations of a sound source. $\hat{\boldsymbol{\phi}}_k$ is statistically related to the corresponding observation vector described in Eq. (1). Both random variables form a special case of a Bayesian network, a Gaussian Mixture Model (GMM)

$$p(\boldsymbol{o}_k|\lambda) = \sum_{i=1}^{M} \pi_i p_i(\boldsymbol{o}_k) \tag{2}$$

composed of M mixture components, with model parameters λ specified as

$$\lambda = \{\pi_i,\, \boldsymbol{\mu}_i,\, \boldsymbol{\Sigma}_i\}, \quad i = 1, \ldots, M.$$

The mixture components in (2) are modeled as D-dimensional, Gaussian distributions $p_i(\boldsymbol{o}_k)$ with mean vectors $\boldsymbol{\mu}_i$, covariance matrices $\boldsymbol{\Sigma}_i$ and mixture weights π_i satisfying $\sum_{i=1}^{M} \pi_i = 1$. Each GMM corresponds to a specific discrete source position in the horizontal plane $\hat{\phi}_{k,1}, \ldots, \hat{\phi}_{k,M}$. In this work, we restricted the number of GMMs to 72, yielding an angular resolution of 5° for the localisation estimates. If new observations are added to the blackboard, the GMMs are triggered to infer the posterior probabilities $p(\hat{\phi}_{k,i}|\boldsymbol{o}_k)$ of all possible locations. The resulting probability distribution $p(\hat{\boldsymbol{\phi}}_k|\boldsymbol{o}_k) = \{p(\hat{\phi}_{k,1}|\boldsymbol{o}_k), \ldots, p(\hat{\phi}_{k,M}|\boldsymbol{o}_k)\}$ is then placed on the blackboard.

To reduce localisation errors caused by front-back confusions, a third layer is introduced in the blackboard architecture that is denoted the *confusion hypothesis layer*. Confusion hypotheses are generated by the *Confusion KS* that operates on this layer. The KS examines if location hypotheses on the second layer contain potential confusions. This examination is based on a threshold $p_{\min} \in [0, 1]$, that defines a

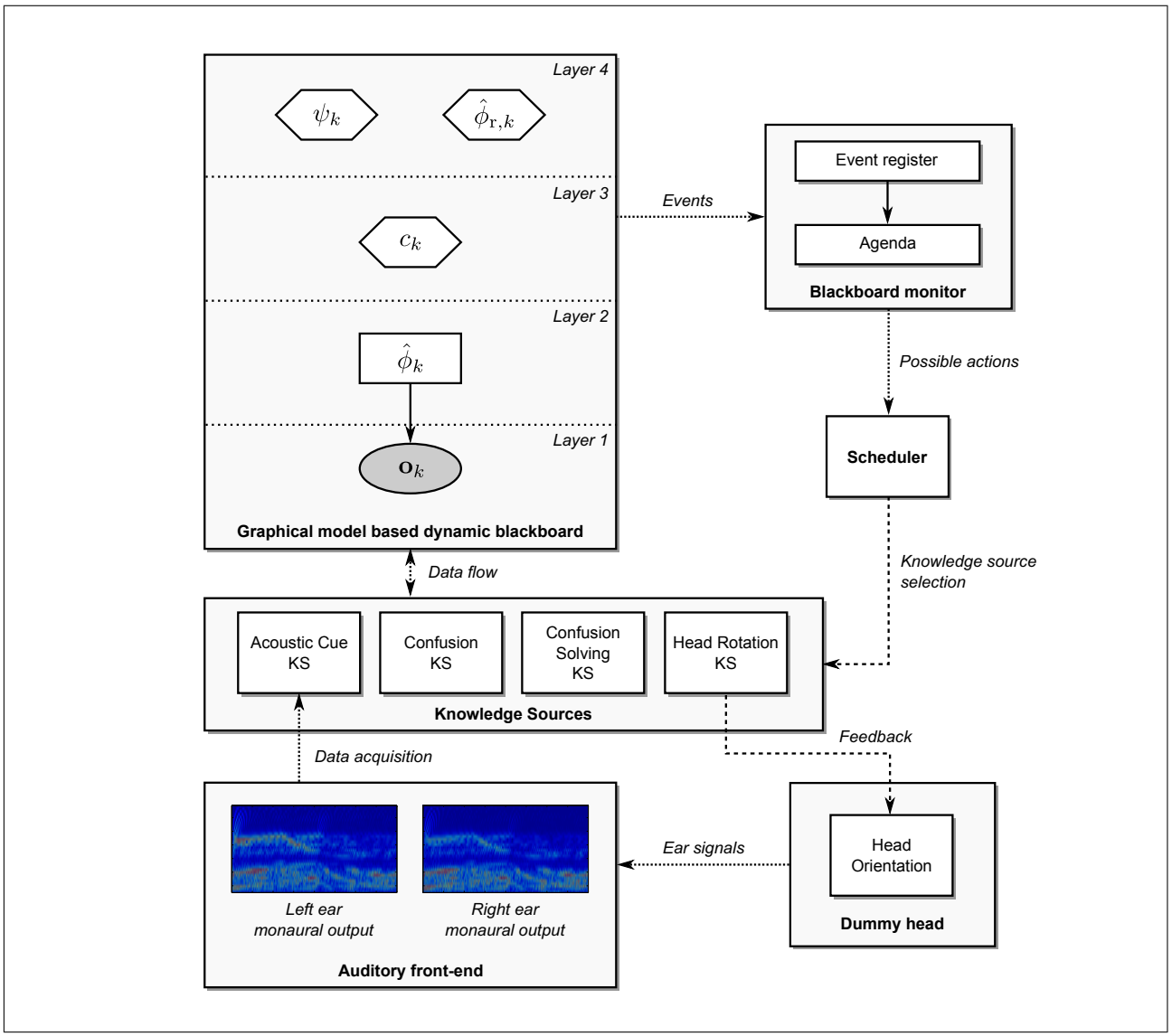

Figure 1. Overview of the proposed blackboard architecture. Data flow between the different components is represented by dotted arrows, whereas dashed arrows represent control commands. The different components on the blackboard are divided into continuous random variables (ellipsoid nodes), discrete random variables (rectangular nodes) and data segments (hexagonal nodes). The GMM that is used in layers 1 and 2 is illustrated by a solid arrow that represents the statistical relationship between the observation vectors $\boldsymbol{o}_k$ and the discrete locations $\hat{\phi}_k$.

probability at which one of the posterior probabilities $p(\hat{\phi}_{k,i}|\boldsymbol{o}_k)$ is considered as a location hypothesis. A confusion is identified if there are multiple location hypotheses within one frame. When a confusion is identified, a confusion hypothesis

$$c_k = \{\tilde{\phi}_{k,1}, \ldots, \tilde{\phi}_{k,Q}\} \tag{3}$$

is created which includes all Q competing locations $\tilde{\phi}_{k,j},\ j = 1, \ldots, Q$. If $Q = 1$, no confusion is detected and a relative source location hypothesis $\hat{\phi}_{\mathrm{r},k}$ is created on the fourth layer of the blackboard.

The fourth layer, denoted as the *perceptual hypotheses layer*, contains two variables ψ_k and $\hat{\phi}_{\mathrm{r},k}$, corresponding to the current head position and the estimated relative source position, respectively. As described before, if no front/back confusion was detected, the estimated relative source position is directly computed by the Confusion KS from the posterior probabilities on the second layer. If there is a remaining confusion hypothesis according to (3) on the third layer and the head has not been rotated, the *Head Rotation KS* is triggered. This halts the listening process and activates the feedback path that triggers a change of the current head orientation. After the rotation is completed, it indicates that the system is ready for the next frame and triggers the *Confusion Solving KS*. This KS solves localisation confusions by predicting the location probability distribution after a head rotation, and comparing it with new location hypotheses that have been gathered within the next

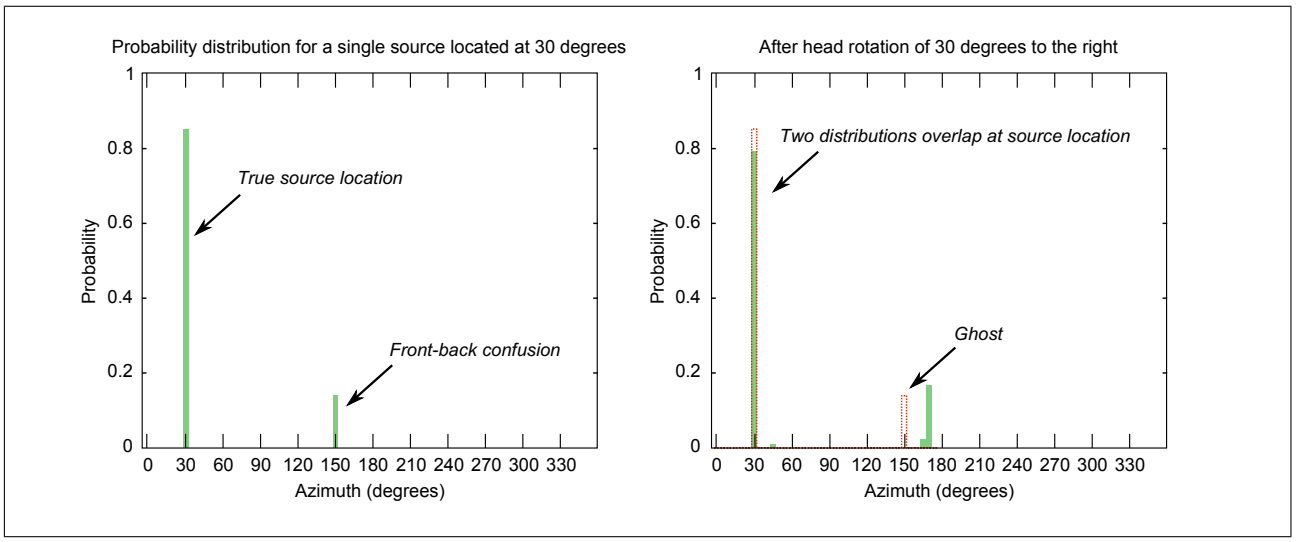

Figure 2. Illustration of front-back confusion solving. The left panel shows the probability distribution for different positions for a source located at 30° azimuth. There clearly exists a *ghost* at 150° azimuth. The right panel shows the predicted location distribution in dotted lines and the actual distribution after head rotation by 10°. The two distributions overlap at 30° azimuth which suggests a true source position.

frame. If a hypothesised source position reflects a true source location, then the predicted location distribution and the observed distribution after head rotation should overlap at the same location. If this is the case, the estimated position is considered a valid relative source location hypothesis $\hat{\phi}_{\mathrm{r},k}$, which is then put onto the blackboard. The corresponding confusion hypotheses on the third layer are then discarded by the Confusion Solving KS. If the predicted and observed distributions do not match, the hypothesised location is considered a *ghost* and the system proceeds with the next frame to gather more data before repeating the process. An example of the confusion solving process is illustrated in Figure 2.

The triggering of specific KSs is attached to certain events that are stored in an event register, which is part of the blackboard monitor. As described before, events are generated if new data is available from the auditory front-end or if specific KSs have performed certain actions on the blackboard. The blackboard monitor keeps track of the current state of the blackboard and generates an agenda which contains all actions that could be performed according to this state.

The agenda is then passed to the scheduler that decides which of the possible actions would be best suited given the current state of the blackboard and the task that should be accomplished. In the current system, a weight is attached to each KS represented as an integer value between 0 and 100. This weight corresponds to the importance of a specific KS for accomplishing the localisation task. Given the agenda, the scheduler executes the action that is linked to the KS with the highest weight.

4. Experiments and results

4.1. Evaluation scenario

The blackboard architecture was evaluated in a single-source localisation scenario. Here the position of the listener was assumed to be static but changes in head orientation were possible. The target sound was a static speech source, but could be located on the horizontal plane at an arbitrary angle between $[0°, 360°]$ with a 5° angular resolution. Since the localisation task was not restricted to the frontal plane, the localisation systems were presented with potential front-back ambiguities.

7 target source positions were selected for evaluation: 270°, 300°, 330°, 0°, 30°, 60°, 90°. Note although the evaluated target source positions were all on the frontal plane, the localisation systems did not have this prior knowledge and assumed an azimuth range of $[0°, 360°]$ for a potential target source position.

Two localisation conditions were evaluated. The first condition contained only the target speech source. The second condition also included a diffuse noise at a signal-to-noise ratio (SNR) of 0 dB in order to evaluate the noise robustness of the proposed system. In both conditions, it was assumed that the listener and the sound source were located in a free-field environment. The simulation of the scenario was generated using HRTFs [14] acquired from a KEMAR dummy-head, recorded at a distance of 3 m between the head and the source.

4.2. Experimental setup

The target source was speech signals taken from the GRID corpus [18]. The GRID corpus consists of short utterances spoken by 34 native English speakers (18

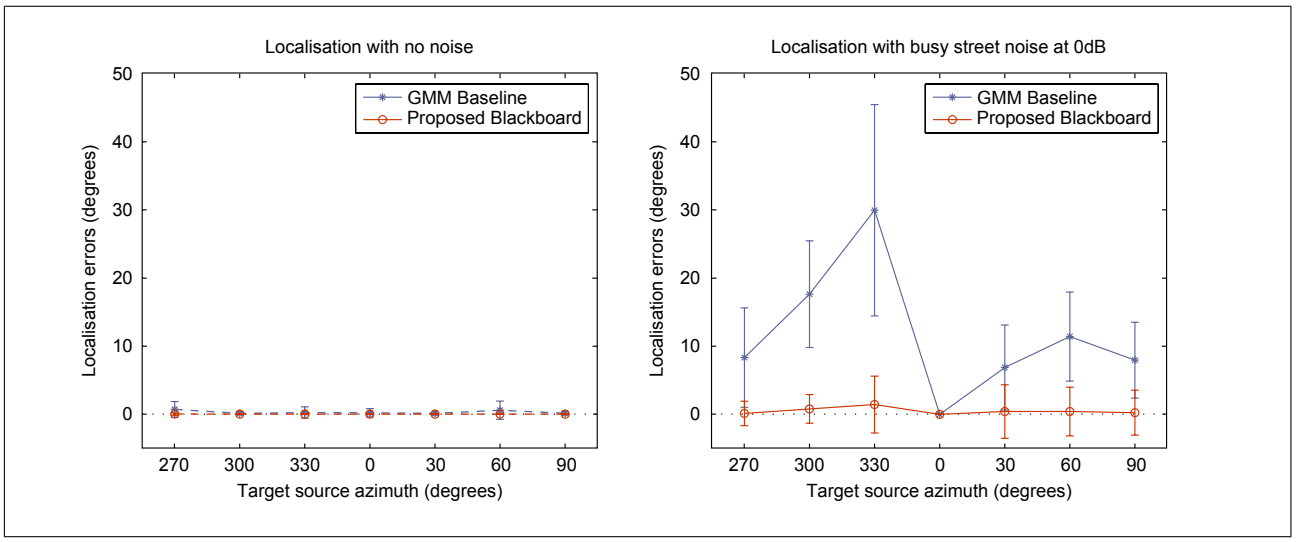

Figure 3. Mean utterance-level localisation errors of the GMM baseline and the proposed blackboard system for localising a speech source. Left: no noise was present. Right: busy street noise was present at an SNR of 0 dB. Error bars show standard deviations.

male speakers and 16 female speakers), in the form <COMMAND> <COLOUR> <PREPOSITION> <LETTER> <NUMBER> <ADVERB>, e.g. "place white at L 3 now". The training set included 340 randomly selected utterances (10 utterances per speaker). They were then spatialised to produce training data for each azimuth position between $[0°, 360°]$ with a 5° step. A further set of 170 utterances (5 utterances per speaker) were selected as the evaluation test set, and were spatialised to simulate the 7 target source positions described above.

The diffuse noise used in the second test condition was one of the environmental sounds ("busy street") taken from IEEE AASP CASA Challenge Dataset [19]. The noise was added to the binaural speech signals after spatialisation at an SNR of 0 dB.

The peripheral processing of the auditory system was simulated by the auditory front-end described in Section 2.2, which decomposed signals arriving at both ears into 31 gammatone filterbank channels. The centre frequencies of the filterbank were equally distributed on the ERB scale between 80 Hz and 8 kHz. The channel output was then halfwave-rectified and used to extract channel-dependent binaural cues. A Hann window of 20 msec was used for analysis in each frame with an overlap between successive frames of 10 msec. The ITD for each channel was estimated by choosing the maximum lag of a cross-correlation function within the range of $[-1, 1]$ msec. The channel ILD was estimated by comparing the energy integrated across the window between the left and right ears within each channel and expressed in dB.

Two localisation systems were evaluated: a GMM-based localisation baseline and the proposed blackboard system. Both systems used GMMs to model the azimuth-dependent distribution of the binaural feature space consisting of ITDs and ILDs. The GMM baseline simply selected the azimuth that has the maximum posterior given a binaural feature observation as the target source position, while the blackboard included top-down feedback for head rotation in order to resolve front/back ambiguities as described in Section 3. To make the two systems more comparable both employed identical sets of GMMs. The GMMs were trained only on spatialised speech signals and no noise was included during training. No prior knowledge of source positions was used.

4.3. Results and discussion

Localisation performance of both systems was evaluated as utterance-level localisation errors. Utterance-level localisation errors were computed by averaging the minimum angular differences between the reference target position and the estimated positions within each utterance. Fig. 3 shows the mean utterance-level localisation errors based on the 170 test utterances for each evaluated target position. Error bars show standard deviations.

In the left panel of Fig. 3 where no noise was present, both systems were able to localise the speech source at all the evaluated positions with very little error. The localisation errors averaged across all target positions were 0.001° and 0.3° for the blackboard system and the GMM baseline, respectively. A t-test showed that performance of the blackboard system was significantly better than that of the GMM baseline ($p < 0.001$). It should be noted that in this clean condition the GMM baseline was able to handle front/back ambiguities without head rotation. This is largely because the GMMs captured the azimuth-dependent patterns of binaural cues across all frequency channels. The subtle spectral difference between front and back was realised by the HRTFs used

in the simulation and thus implicitly modelled by the system.

When diffuse noise was present, shown in the right panel of Fig. 3, the localisation errors of the GMM baseline increased significantly across all target positions except for the 0° azimuth (average localisation errors across all target positions: 11.8°). Performance was particularly bad for the GMM baseline at azimuth positions where the front-back confusion was strong (30° and 60° at both sides). The performance of the blackboard system, however, was generally robust in the presence of the diffuse noise (average localisation errors across all target positions: 0.5°) and was significantly better than the baseline (t-test; $p < 0.001$). The top-down feedback that allowed head rotation helped the system resolve most ambiguities and the improvement over the baseline was consistent across all the target positions.

5. Conclusions and future work

We have presented a general high-level framework for auditory scene analysis, which, based on a graphical-model representation, can iteratively develop an "understanding" — an internal, interpretable high-level description — of an auditory scene. While results were shown for a small toy example, consisting of localisation of a single acoustic source, the framework allows inference in a wide range of dynamic Bayesian networks, supporting many types of knowledge sources and inference strategies.

Thus, a natural next step will be the integration of dynamic state-space models, describing sources not as stationary but as dynamically moving. Tracking these dynamic sources will hence become necessary. In the proposed framework, this can be achieved by incorporating a source-type-dependent state-space model for the source position. Inference of the source position will hence be possible by developing Kalman-style filters. These, based on strategies like the unscented Kalman filter [20], can additionally include the blackboard's estimates of the uncertainties of all graphical-model variables to obtain optimal, time-varying estimates of all source positions.

To test feedback strategies in a controlled environment, we plan to integrate our blackboard architecture with the *Bochum Experimental Feedback Testbed* (BEFT) [21], a tool that has been designed to test complex feedback strategies early in the TWO!EARS project. BEFT provides a custom-made virtual environment for visualization of XML-scripted scenes and allows the success of actual feedback strategies to be monitored in near real-time. Visual and (emulated) auditory features provided by the BEFT system core will act as input to our KSs. To that end, the testbed architecture is explicitly designed to closely approximate real-life conditions: ground-truth characteristics of environmental objects are artificially degraded to mimic weak sensor performance under adverse environmental conditions. The BEFT framework is not limited to the emulation/degradation of physical object or scenario features: specific degradation functions allow emulating an object classifier that provides category labels for each observed environmental entity and generates input to higher-level KSs. Further, the testbed architecture provides task stack mechanisms to control the behavior of a virtual robotic platform that explores a given scenario.

In conclusion we have shown that the extension of machine hearing systems with top-down feedback prove to be advantageous over those that are solely based on bottom-up processing. Graphical-model-based blackboard systems, as introduced in this work, are a flexible framework for further investigating the role of feedback in the context of machine hearing systems. Focusing on human perception, the blackboard paradigm furthermore allows for easy integration of additional cues like visual and tactile information, providing a powerful framework for biologically inspired computational systems.

Acknowledgement

This research has been supported by EU FET grant TWO!EARS[1], ICT-618075. We thank Tobias May for making the auditory front-end code available.

References

[1] A. S. Bregman: Auditory scene analysis: the perceptual organization of sound. Cambridge, MA: MIT Press, 1990.

[2] J. Blauert: Spatial hearing – The Psychophysics of Human Sound Localization. Cambridge, MA: MIT Press, 1997.

[3] D. Wang, G. J. Brown (Eds.): Computational auditory scene analysis: Principles, Algorithms, and Applications. IEEE Press/Wiley-Interscience.

[4] H. Wallach: The role of head movements and vestibular and visual cues in sound localization. J. Exp. Psychol. 27(4):339–368, 1940.

[5] V. A. F. Lamme, H. SupÃÍr, H. Spekreijse: Feedforward, horizontal, and feedback processing in the visual cortex. Current Opinion in Neurobiology 8(4):529–535, 1998.

[6] B. R. Schofield: Structural organization of the descending auditory pathway, in: Oxford Handb. of Auditory Science, Vol. 2: The Auditory Brain. Oxford Univ. Press, New York, NY, 2009.

[7] A. Rabiee, S. Setayeshi, S. Y. Lee: CASA: Biologically Inspired Approaches for Auditory Scene Analysis. Natural Intelligence: the INNS Magazine 2(1):50–58, 2012.

[8] L. D. Erman, F. Hayes-Roth, V. R. Lesser, D. R. Reddy: The Hearsay-II speech understanding system: Integrating knowledge to resolve uncertainty. ACM Comput. Surv. 12(2):213–253, 1980.

[1] http://www.twoears.eu

[9] J. Pearl: Bayesian Networks: A Model of Self-Activated Memory for Evidential Reasoning. Proceedings of the 7th Conference of the Cognitive Science Society, University of California, Irvine, CA. pp. 329âĂŞ-334, 1985.

[10] M. Cooke, G. J. Brown, M. Crawford, P. Green: Computational auditory scene analysis: listening to several things at once. Endeavour, 4, 186–190, 1993.

[11] V. R. Lesser, S. H. Nawab, F. I. Klassner: IPUS: An architecture for the integrated processing and understanding of signals. Artificial Intelligence, 77, 129–171.

[12] D. Ellis: Prediction-driven computational auditory scene analysis. PhD Thesis, MIT, 1996.

[13] D. Godsmark, G. J. Brown: A Blackboard Architecture for Computational Auditory Scene Analysis. Speech Communication, 27, 351–366, 1999.

[14] H. Wierstorf, M. Geier, A. Raake, S. Spors: A Free Database of Head-Related Impulse Response Measurements in the Horizontal Plane with Multiple Distances. 130th Convention of the Audio Engineering Society, 2011.

[15] T. May, S. van de Par, Armin Kohlrausch: A Probabilistic Model for Robust Localization Based on a Binaural Auditory Front-End. IEEE Transactions on Audio, Speech, and Language Processing 19(1):1–13, 2011.

[16] B. R. Glasberg, B. C. J. Moore: Derivation of auditory filter shapes from notched-noise data. Hearing research 47(1):103–138, 1990.

[17] B. C. J. Moore, B. R. Glasberg, T. Baer: A model for the prediction of thresholds, loudness and partial loudness. J. Audio Eng. Soc., vol. 45, pp. 224–240, 1997.

[18] M. Cooke, J. Barker, S. Cunningham, X. Shao: An audio-visual corpus for speech perception and automatic speech recognition. Journal of the Acoustical Society of America, 120, 12421–24, 2006.

[19] D. Giannoulis, E. Benetos, D. Stowell, M. D. Plumbley: IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events - Public Dataset for Scene Classification Task, Queen Mary University of London, 2012.

[20] S. J. Julier, J. K. Uhlmann: A new extension of the Kalman filter to nonlinear systems. Int. Symp. Aerospace/Defense Sensing, Simul. and Controls 3: 182, 1997

[21] T. Walther, B. Cohen-Lhyver: Multimodal feedback in auditory-based active scene exploration. EAA Forum Acusticum, KrakÃşw, 2014 (submitted)