Unsupervised Inference of Auditory Attention from Biosensors

Melih Kandemir[1], Arto Klami[1], Akos Vetek[2], and Samuel Kaski[1,3]

[1] Helsinki Institute for Information Technology HIIT
Department of Information and Computer Science
Aalto University
[2] Nokia Research Center
Media Technologies Lab
[3] Helsinki Institute for Information Technology HIIT
Department of Computer Science
University of Helsinki

Abstract. We study ways of automatically inferring the level of attention a user is paying to auditory content, with applications for example in automatic podcast highlighting and auto-pause, as well as in a selection mechanism in auditory interfaces. In particular, we demonstrate how the level of attention can be inferred in an unsupervised fashion, without requiring any labeled training data. The approach is based on measuring the (generalized) correlation or synchrony between the auditory content and physiological signals reflecting the state of the user. We hypothesize that the synchrony is higher when the user is paying attention to the content, and show empirically that the level of attention can indeed be inferred based on the correlation. In particular, we demonstrate that the novel method of time-varying Bayesian canonical correlation analysis gives unsupervised prediction accuracy comparable to having trained a supervised Gaussian process regression with labeled training data recorded from other users.

Keywords: Affective computing, Auditory attention, Canonical correlation analysis.

1 Introduction

Attention to external stimulation is a central element in human cognition. By selectively focusing on specific aspects of the stimulation we can control the information gain, to maximally utilize the limited information channels. In Human-Computer Interaction (HCI), attention plays several roles: Information in the user interface should be structured to capture users attention by making it salient when it needs attention [22], but it is also possible to use the attention of the user as a form of implicit input. For visual attention, eye-tracking devices provide a direct interface for measuring attention; they have been used in a range of attentive interfaces, starting from eye tracking based zooming of

P. Flach et al. (Eds.): ECML PKDD 2012, Part II, LNCS 7524, pp. 403–418, 2012.

windows (for a review of attentive interfaces see [25]) to using eye tracking for estimating aspects such as topical relevance in information retrieval [15,20].

Here we venture beyond visual attention to auditory attention. For vision the eye-tracking devices provide relatively direct access to the target of the attention, which has enabled the extensive works on utilizing the attention target as part of the interface design. For auditory attention, however, detecting even where the user is paying attention is largely an open issue, and no simple hardware solutions exist for recording it. In particular, the best current methods are based on direct recording of neuronal activity using functional MRI [16,19] and MEG [12], which are by no means feasible for human-computer interaction, or full-scalp EEG (see [11] for an early example) which is also impractical. For a good overview of auditory attention and extensive list of references, see [7].

In this work we will discuss machine learning approaches useful for creating more portable auditory attention detection devices. Due to the general difficulty of the task, we will consider the simplified task of estimating *how much* a person is attending to particular auditory content. The approach could be directly generalized to the task of estimating to which of multiple parallel auditory streams the user is focusing on, by comparing the level of attention paid to each of the streams, but to simplify the experimental setup we consider explicitly only the task of measuring the amount of attention for a single source.

While specific hardware focusing on auditory attention is lacking, we revert to the choice of using a combination of available physiological sensors for recording the state of the user. We record neuronal activity with an easy-to-wear single-channel EEG, the amount of body movement with an accelerometer, and eye movements with an eye-tracker. While these sensors are clearly not optimal for detecting auditory attention, they still provide multivariate signals that represent the activity of the individual user while she is listening to some auditory content. The field of affective computing studies the use of such signals for inferring various cognitive and affective properties of the user, and relatively good success has been demonstrated for instance in inferring emotional valence and arousal [5,18], specific emotions [13], and mental workload [28]. Hence, it is a reasonable assumption that we could get a handle on the attention with similar sensors as well. In fact, [17] has already demonstrated success in discovering loss of auditory attention due to external interruptions by monitoring the galvanic skin response.

Given the sensory signals, the task of detecting auditory attention is, in principle, a straightforward learning task. We merely need to obtain ground truth training labels and train a classifier or a regression model for inferring the labels from the signals. For a classifier, the labels would be high vs. low attention, and for a regression model the actual level of attention. However, it is extremely challenging to collect training labels for the task of inferring the level of attention. For example, if the subject is listening to a music piece, we cannot ask him to continuously rate his level of attention since providing that feedback would change his behavior; needing to provide the evaluation would prevent him from naturally attending to the music. It is also unreasonable to expect that people would be able to quantify their level of attention to arbitrary auditory contents after the experiment with high accuracy.

The only remaining way of collecting the training labels is to conduct a laboratory experiment where the labels stem from a controlled experiment. In this work, we present a simple experiment of that sort, using an additional visual task of varying complexity to control the level of attention remaining for an auditory listening task. While such a procedure gives training labels, it is important to realize that it will only provide them for the users who took part in the particular laboratory experiment; it still remains infeasible to obtain any training data whatsoever for the eventual users of an auditory attention detector. This observation implies that any model inferring the level of attention from the sensory signals must be user-independent. Such models, in turn, are known to be of relatively poor accuracy due to considerable user-specific variation in the sensory signals. Nevertheless, in this work we demonstrate that we can infer the level of auditory attention with reasonable accuracy using user-independent supervised models, by applying two state-of-art probabilistic kernel-based regressors: Gaussian process regression [21] and Relevance Vector Machines [23].

Our main contribution, however, is an alternative way of inferring the level of attention that does not require any training labels whatsoever. Instead, we make a hypothesis that the amount of synchrony or correlation between the physiological signals and the auditory content is modulated by the level of attention. That is, we assume that any signals recorded from a user not paying attention to the audio will be independent of the audio signal, whereas high degree of attention is reflected as increased correlation between some of the physiological signals and the audio content. Assuming the hypothesis holds, we can directly detect auditory attention as correlation between the two signals, without needing any training data. To further illustrate the approach and its relationship with the supervised one, the analysis pipelines for both are depicted in Figure 1.

We measure the correlation with canonical correlation analysis (CCA) and its Bayesian re-formulation as a latent variable model [2,27]. Using the Bayesian formulation not only helps with limited amount of data, but enables encoding prior knowledge on the underlying signals into the model. In this work we utilize the fact that the measurements are time series, and introduce a novel time-dependent Bayesian CCA model by encoding time-dependent interactions in the generative description. We learn the model from the coupled physiological signals and features computed for the audio content, and then measure the amount of correlation to represent the level of attention. We demonstrate that the correlation reveals the level of attention with accuracy comparable to the user-independent supervised models. The empirical experiments hence demonstrate that we can infer the level of attention from physiological signals, and more importantly that we can do it without requiring any labeled training data at all.

We start the rest of the paper by first introducing some prototypical application scenarios for auditory attention detection, providing a context and motivation for the more technical sections. We then proceed to explain the computational models needed both for supervised user-independent inference of attention and for

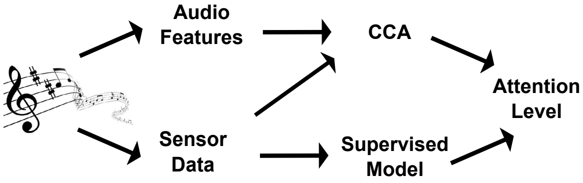


Fig. 1. Illustration of the two modeling pipelines for estimating the level of attention from the biosensors. The unsupervised approach evaluates the correlation between the sensor data and the audio content, and uses the inverse correlation as a predictor for the level of attention. The supervised approach uses only the sensor data, and applies a model learned from other users for predicting the level of attention.

measuring the amount of correlation between the physiological signals and the auditory content. After describing the models we explain the empirical experiment conducted for recording data to train and evaluate the models, and then show the empirical results demonstrating the accuracy of the proposed methods.

2 Application Overview

Albeit we here consider the task of inferring the level of auditory attention primarily as a basic research question, it has several direct application possibilities that are worth highlighting. A simple example would be an auto-pause tool for audio players; the attention recognizer would be running continuously on the background and whenever it recognizes that the level of attention is particularly low it pauses the audio automatically so that the user can continue listening for the audio after the interfering concurrent task is over. Alternatively, the tool can simply keep track of the moments with low attention, allowing the user to easily return to them later (for a practical example, see [17]), for example to re-cap details of a technical description in a podcast they might have missed during the first listening.

There are also applications where storing the moments with the *highest* attention could be useful. Such an automatic highlighting tool could capture, for example, the moments when the user most enjoyed a piece of music. Those pieces could even serve as a query to a music retrieval engine; the user could carry out a search for other songs similar to the most enjoyable parts of the song he just listened. Besides static content, such as music or podcasts, the tool could also be used for highlighting more dynamic content. For example, it could be used to summarize a meeting as a combination of the moments where reasonable amount of attention was paid to the discussion.

In another application scenario the goal is to detect the attention target. In an environment with multiple overlapping auditory streams, we can measure the amount of correlation with respect to each of the sources and detect the target of primary attention as the one with the highest correlation. This enables building for instance auditory interfaces where attention is used to implicitly select one out of multiple alternatives.

3 Modeling Dependencies

We first describe the classical method for estimating the amount of multivariate correlation between two data sources, the canonical correlation analysis (CCA). We then proceed to describe the Bayesian variant of CCA, following the formulation in [27], which makes CCA applicable for high-dimensional data and allows various extensions. Finally, we introduce the novel time-dependent Bayesian CCA model that explicitly models continuity in time-series data.

3.1 CCA

Given two data matrices, $\mathbf{X} \in \mathbb{R}^{N \times D_x}$ and $\mathbf{Y} \in \mathbb{R}^{N \times D_y}$, with N samples (rows) and D_x and D_y features (columns), the CCA finds linear projection weights $\mathbf{u} \in \mathbb{R}^{1 \times D_x}$ and $\mathbf{v} \in \mathbb{R}^{1 \times D_y}$ such that the Pearson correlation

$$\rho = \mathrm{cor}(\mathbf{X}\mathbf{u}^T, \mathbf{Y}\mathbf{v}^T)$$

is maximized. Since correlation is invariant to the scale, the norms of $\mathbf{u}$ and $\mathbf{v}$ can be fixed to unity. The above formulation defines the most correlating one-dimensional subspace; further components indexed by a subscript can be obtained by adding an orthogonality constraint $cor(\mathbf{X}\mathbf{u}_k^T, \mathbf{X}\mathbf{u}_l^T) = 0$ for all k and l (and similarly for $\mathbf{Y}$). In practice, we can readily compute $\min(D_x, D_y)$ canonical correlations ρ_k and the associated projections $(\mathbf{u}_k, \mathbf{v}_k)$ by solving a single generalized eigenvalue problem (see for, instance, [10] for details).

While CCA is typically used for gaining an understanding of the correlations between the two data sets (by interpreting $\mathbf{u}$ and $\mathbf{v}$), it readily provides a measure for the *amount of dependency* between them. We summarize the dependency with the quantity

$$I(\mathbf{X}, \mathbf{Y}) = -\frac{1}{2} \sum_{k=1}^{\min(D_x, D_y)} \log\left(1 - \rho_k^2\right),$$

which corresponds to the mutual information between the two sources if they are jointly multivariate normal. For non-normal data the quantity does not correspond to the mutual information, but is still a good estimate of the total dependency, summarizing all correlations into a single number.

In our application, the task is to measure the amount of correlation for a subset of the samples. We do this by first learning the model to maximize the correlation over the whole available data. Given the model (the projections), we then evaluate the correlation for any subset L of the samples by simply estimating the correlations (and the above mutual information summary) between $\mathbf{X}_L\mathbf{u}_k^T$ and $\mathbf{Y}_L\mathbf{v}_k^T$.

3.2 Bayesian CCA (BCCA)

While CCA is a straightforward method with guaranteed convergence to a global optimum, it has a number of shortcomings that can be addressed by switching to the probabilistic framework. First of all, CCA is prone to overfitting to high-dimensional data, and especially for $N < \min(D_x, D_y)$ necessarily returns correlations of exactly one due to linear dependency of the features. For preventing this we need proper regularization, which we implement through priors and variational inference in the Bayesian framework. This particular choice gives us also another advantage: It allows extending the model by making slight changes to the generative model of CCA (see for instance [1,8,26] for examples). After recapping the model, we will in the next section show how the Bayesian treatment of CCA enables incorporating time-dependencies between the samples, a property that would not be easy to add in the original linear algebraic formulation.

The Bayesian CCA builds upon the probabilistic interpretation of CCA by [2]. The basic idea is that the two data sources are generated from a common latent representation with a linear transformation, with arbitrary additive Gaussian noise independent of the other source. More formally,

$$\begin{aligned} \mathbf{z} &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \\ \mathbf{x} &\sim \mathcal{N}(\mathbf{W}_x \mathbf{z}, \boldsymbol{\Psi}_x) \\ \mathbf{y} &\sim \mathcal{N}(\mathbf{W}_y \mathbf{z}, \boldsymbol{\Psi}_y), \end{aligned} \tag{1}$$

where $\mathbf{z} \in \mathbb{R}^{1 \times K}$ is a K-dimensional latent signal and $\mathbf{W}_x \in \mathbb{R}^{D_x \times K}$ and $\mathbf{W}_y \in \mathbb{R}^{D_y \times K}$ are projections mapping the latent signals to the observations. The noise covariances $\boldsymbol{\Psi}_x$ and $\boldsymbol{\Psi}_y$ model the variation independent of the other sources. In practice, especially for high-dimensional data, we need to assume low-rank noise covariances $\boldsymbol{\Psi}_x$ and $\boldsymbol{\Psi}_y$ to prevent needing to make inference over the $D_x \times D_x$ and $D_y \times D_y$ covariance matrices, which leads to the Bayesian CCA formulation of [27]:

$$\begin{aligned} \mathbf{z} &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \\ [\mathbf{x}; \mathbf{y}] &\sim \mathcal{N}(\mathbf{W}\mathbf{z}, \boldsymbol{\Sigma}), \end{aligned}$$

where $\boldsymbol{\Sigma}$ is a block-diagonal matrix $\boldsymbol{\Sigma} = [\sigma_x^2 \mathbf{I}, \mathbf{0}; \mathbf{0}, \sigma_y^2 \mathbf{I}]$. By making $\mathbf{W}$ group-wise sparse with the sparsity-inducing prior

$$\begin{aligned} \beta_{xk} &\sim \mathcal{G}(\alpha_0, \beta_0) \\ \beta_{yk} &\sim \mathcal{G}(\alpha_0, \beta_0) \\ p(\mathbf{W}) &= \prod_{k=1}^{K} \left(\mathcal{N}(\mathbf{W}_x(k) | \mathbf{0}, \beta_{xk}^{-1} I) \mathcal{N}(\mathbf{W}_y(k) | \mathbf{0}, \beta_{yk}^{-1} I) \right), \end{aligned}$$

where $\mathcal{G}(\alpha_0, \beta_0)$ is a flat Gamma distribution ($\alpha_0 = \beta_0 = 10^{-14}$), we will get projections $\mathbf{W}$ that factorize as

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_x & \mathbf{V}_x & \mathbf{0} \\ \mathbf{W}_y & \mathbf{0} & \mathbf{V}_y \end{bmatrix}.$$

After marginalizing out the latent components corresponding to the columns of $\mathbf{W}$ having a zero block for either data source, induced by the group-wise sparsity prior, the above model becomes equivalent to (1) with $\boldsymbol{\Psi}_X = \mathbf{V}_x\mathbf{V}_X^T + \sigma_x^2\mathbf{I}$ and $\boldsymbol{\Psi}_y = \mathbf{V}_y\mathbf{V}_y^T + \sigma_y^2\mathbf{I}$. In summary, the resulting model implements the Bayesian CCA model with a low-rank assumption for the noise covariances within each data source, but does not require specifying the rank of either the correlating subspace or the noise covariances in advance. Instead, they will all be learned automatically from the data.

We do the inference using a variational approximation, assuming the posterior $p(\Theta|\mathbf{X}, \mathbf{Y}) = p(\sigma_x^2, \sigma_y^2, \beta_x, \beta_y, \mathbf{Z}, \mathbf{W}|\mathbf{X}, \mathbf{Y})$ can be approximated by a factorized distribution

$$Q(\Theta) = q(\sigma_x^2)q(\sigma_y^2)\prod_{k=1}^{K} q(\beta_{xk})q(\beta_{yk})\prod_{i=1}^{N} q(\mathbf{z}_i)\prod_{d=1}^{D} q(\mathbf{W}_{d\cdot}),$$

and minimizing the Kullback-Leibler divergence between $Q(\Theta)$ and $p(\Theta|\mathbf{X}, \mathbf{Y})$. This results in a set of mean-field equations updating each term $q(\cdot)$ at a time until convergence to a local optimum; the details can be found in [27].

For evaluating the correlation we estimate the conditional densities $p(\mathbf{z}|\mathbf{x})$ and $p(\mathbf{z}|\mathbf{y})$ and then compute the correlation between their expectations. This can be done for any subset of the samples using a procedure similar to updating $q(\mathbf{Z})$ while learning the posterior approximation.

3.3 Time-Dependent Bayesian CCA (T-BCCA)

One advantage of the Bayesian formulation for CCA is that it allows easily extending the model to take into account particular properties of the underlying data. As practical examples, in [1,26] more robust variants were introduced by replacing the normal distributions with t-distributions, in [14,26] mixtures of CCAs, and in [8] sparse variants. In this paper, we will extend Bayesian CCA to a state-space model that is more accurate for modeling correlations between two multivariate time series.

The key idea of the novel time-dependent CCA is that the latent variables z will have a Markovian assumption. Instead of drawing each z_t independently from the same prior, we introduce the prior

$$\begin{aligned}\mathbf{z}_0 &\sim N(\mathbf{0}, \mathbf{I})\\ \mathbf{z}_t &\sim \mathbf{T}\mathbf{z}_{t-1} + N(\mathbf{0}, \sigma_0^2\mathbf{I})\end{aligned}$$

where $\mathbf{T}$ governs the evolution of the latent space and σ_0^2 controls the amount of stochastic noise.

We retain the variational Bayesian framework for inference, and are able to reuse the update formulas for the various terms in the approximation except for $q(\mathbf{Z})$. For that, we learned a Kalman filter along with the Rauch-Tung-Striebel smoother [9], using a forward-backward procedure as described in [3]. Since the Bayesian CCA assumes independent latent components, we restrict $\mathbf{T}$ to be diagonal to avoid modeling dependencies between them and use variational inference with prior centered around the identity matrix. Furthemore, we set $\sigma_0^2 = 1$ to fix the scale, but could do variational inference over this parameter as well.

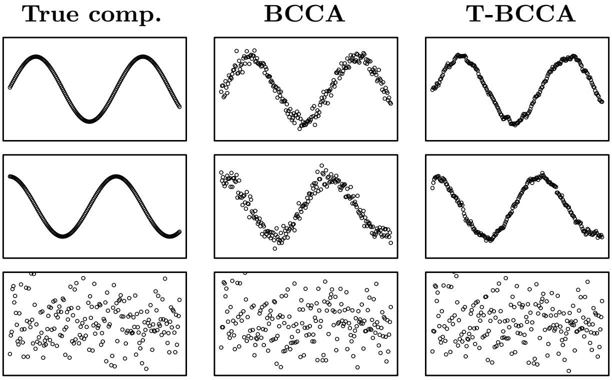


Fig. 2. Illustration of the importance of modeling time-dependencies in CCA modeling. The left column shows three latent components underlying a generated data set, the middle column shows the latent components estimated by regular Bayesian CCA, and the right column shows the estimated components when modeling also the time-dependency. We see that already the regular Bayesian CCA (BCCA) captures the signals roughly correctly, but that the time-dependent model (T-BCCA) gives much more accurate estimates for the sinusoidal signals. The component with no time-dependencies (bottom row) is modeled equally well; T-BCCA learns to automatically set the corresponding element in $\mathbf{T}$ to zero.

To briefly illustrate the advantage of modeling time-dependencies in the latent space, we applied both the regular Bayesian CCA model described in the previous section (which corresponds to $\mathbf{T} = 0$ and $\sigma_0^2 = 1$ in the more general formulation) and the time-dependent model to a simple simulated data. The data has three latent components of which some show clear time-continuity, and as shown in Figure 2 modeling the time-series nature results in more accurate estimates for them.

4 Supervised Learning

Given the available sensor data, the usual approach for inferring the level of attention would be to use supervised learning. That is, we would collect labels for training instances and train a classifier or regressor for predicting the attention. As discussed in the Introduction, gathering such labeling data for the task of auditory attention is extremely challenging and the only reasonable way is to use laboratory experiments with controlled stimulation. This allows gathering training data from laboratory users, but not from the eventual users of an auditory attention predictor system. Hence we require user-independent models in this task.

In this paper, we will compare the unsupervised approach with supervised learning where the model is learned from a training corpus measured and labeled for other users. We chose two state-of-art supervised methods: Relevance Vector Machine (RVM) [23] and Gaussian process (GP) regression [21], as representative alternatives. Both are probabilistic kernel-based learning methods that enable non-linear mappings from the input to the level of attention. To capture non-linearity, we used the Radial Basis Function (RBF) kernel:

$$k(\mathbf{x}, \mathbf{x'}) = \exp(\frac{\|\mathbf{x} - \mathbf{x'}\|^2}{2\nu^2})$$

with both of these methods. For each method, we learned the kernel parameter ν using type II maximum likelihood.

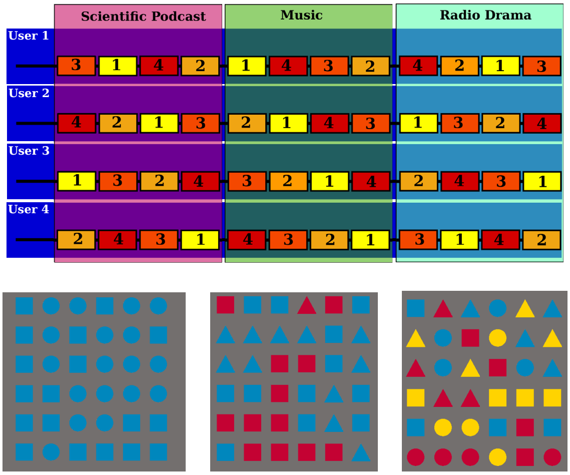
5 Experimental Setup and Data

To train and evaluate the proposed methodology for inferring the level of auditory attention, we created a simple laboratory experiment. The subjects listen to three types of audio content (scientific podcast, popular music, and audio drama) while being measured with three different sensors (NeuroSky single-channel EEG device, accelerometer measuring body movement, and eye-tracker measuring the pupil dilation). Their level of attention to the auditory content is controlled by a simultaneous alternative task with tunable difficulty competing for their attentional resources. Based on the limited-capacity theorem asserting that there is a direct performance tradeoff between simultaneous auditory and visual tasks [4], we assume that the auditory attention is low whenever the user is paying a high level of attention to the alternative task, and vice versa.

For the alternative task we chose a visual search task called *conjunction search* where the user searches for objects identified by multiple features [24]. The user is presented a grid of items, and asked to tell whether any of the items on the screen is unique in terms of color and shape. We assigned the user the binary detection task, instead of asking him to point where the unique item is, to avoid introducing unnecessary movements.

The visual task was presented in four difficulty levels. We tuned the difficulty of the search task by the number of different colors and shapes of the objects. The easiest level (level 1) was simply the blank screen, hence there was no search task at all. The remaining three levels of difficulty had 2, 4, and 9 different kinds of objects, respectively (see the bottom half of Figure 3 for illustration of the stimuli). This provides data with four ground-truth levels of visual attention, and we assume that the auditory attention has an inverse monotonous relationship to visual attention.

We constructed a partially balanced experimental setup for our 12 voluntary test users (7 male and 5 female university students aged from 22 to 29 years). All of them listened to the three audio contents, a scientific podcast, music, and radio drama, of 4 minutes each. There was 1 minute of each visual task level within each audio type. The order of the visual task levels within each audio type was balanced across users using the 4 × 4 Latin squares design. The order



	Scientific Podcast				Music				Radio Drama			
User 1	3	1	4	2	1	4	3	2	4	2	1	3
User 2	4	2	1	3	2	1	4	3	1	3	2	4
User 3	1	3	2	4	3	2	1	4	2	4	3	1
User 4	2	4	3	1	4	3	2	1	3	1	4	2

Fig. 3. Top: Illustration of the experimental setup. The subjects listened to three types of audio content while performing visual search tasks of varying difficulty (four levels). The blocks of the visual tasks occurred at the same time for all subjects, but the design was balanced so that the difficulty levels were in a different order for different subjects, as well as in a different order between the different audio types for each subject. Furthermore, the subjects listened the three audio types in different orders (not shown in the image for clarity). **Bottom:** Examples of the three visual search task difficulties corresponding to levels 2, 3, and 4 from left to right. Level 1 is blank screen where there is no visual search task at all.

of the audio contents was also balanced according to another 3×3 Latin squares. Figure 3 illustrates the course of the experiment.

For the data analysis we processed both the auditory content and the biosignals into vectorial samples, forming each sample from a 250ms contiguous block of the signals. The music is represented by 17 numerical features capturing primarily timbral and rhytmic properties of the music, computed using the MIR toolbox [6]. The idea is that the representation would capture essential characteristics of the audio content. The physiological signals, in turn, were summarized through several features stemming from the affective computing literature, considered to be reasonable approximations of the information content in the physiological signals. The actual features used are listed in Table 1.

6 Results

For evaluating the accuracy and feasibility of the proposed attention inference method, we ran two separate computational experiments on the experimental data described in the previous section. Here we both describe the experiments and report their results.

Table 1. Top: Physiological features extracted from the data collected by three sensors: accelerometer, eye-tracker, and single-channel EEG. **Bottom:** Audio features computed by the MIR toolbox [6].

Physiological features
3D body motion and pupil diameter: mean and standard deviation mean of the derivative, mean, median, and maximum peak-to-peak interval
Single-channel EEG: spectral power in (0.5–2.75) Hz, (3.5–6.75) Hz, (7.5–9.20) Hz,(10.0–11.75) Hz, (13.0–16.75) Hz,(18.0–29.75) Hz, (31.0–39.75) Hz, (41.0–49.75) Hz

Audio features
zero-crossing rate, spectral centroid, brightness, spectral spread, kurtosis, MFCC (Mel-frequency cepstral coefficients), skewness, roll-off, entropy, spectral-flatness, roughness, RMS (root-mean-square), spectral flux, novelty of spectral flux, fluctuation, fluctuation centroid, fluctuation entropy

6.1 Experiment 1: Inferring the Level of Attention for Long Time Blocks

In the first experiment we study the problem of inferring the level of attention for each of the experimental blocks. This answers the question of whether we can differentiate between different levels of attention during periods of time lasting roughly one minute each. The results would be directly applicable to scenarios such as meeting highlighting but would not be sufficient for choosing the attention target in an interactive interface, for instance.

We analyze each of the audio types and users separately, resulting in a total of $12 \times 3 = 36$ models. For the correlation-based models we learn the CCA using all the data for that user-audio pair and then evaluate the correlation for each of the blocks corresponding to one level of ground truth attention. Since we have four levels of ground truth attention, this gives us four correlation scores which we sort in the decreasing order to predict the attention. For each user-audio pair we then compute the accuracy as the number of correct ranks with respect to the ground truth. For example, if the block with the hardest visual task is ranked last, the score increases by one. This measure is equivalent to classification accuracy for a scenario where we know that each class occurs only once in the test set.

For the supervised models we train a regression model with the labeled training data for all other users and then apply it to the four blocks of the user in question. This is done separately for each audio type, and we again rank the resulting regression scores to label the four blocks with the levels of attention. That is, we use the exactly same measure as for the correlation-based models.

Table 2 collects the average scores (over users, normalized so that 1 equals to a perfect result) for all of the methods and all three audio types. The supervised user-independent approaches provide the best results, but the unsupervised variants closely follow with only marginally lower accuracies. Of these three, the

T-BCCA model has an accuracy over 40% for all three audio types, which is a promising signal for real applications. To evaluate the reliability of the results we performed Wilcoxon signed rank test over the results of all three audio types (to obtain more statistical evidence over the $12 \times 3 = 36$ independent scores), revealing that all five methods are significantly ($p < 0.05$) better than the random baseline, but that the differences between the alternatives are not significant.

Note that in this experiment we normalized the correlations by dividing them by the mean of the correlations for all users during the same audio content. This was done to reduce the potential bias caused by the properties of the auditory content itself. It is easy to imagine that certain types of audio content, for example catchy music pieces, could result in naturally higher correlation levels with the sensory signals for all attention levels. The kind of normalization done in this experiment could be done for real applications given access to sensory data of other users having listened to the same content, still without needing any labeled data. As this is not necessarily the case in many situations, we also re-ran the experiment without such normalization. This results in a drop of (on average) two percentage points for each of the unsupervised methods. Together these two experiments indicate that it pays off to remove the content-specific effect on the correlation, but that it is not absolutely crucial and the methods work even without any earlier data from other users.

Table 2. The classification accuracy for detecting the four levels of attention in the experiment. All five methods outperform the chance level with statistical significance ($p < 0.05$; Wilcoxon). The noteworthy observation is that the unsupervised CCA-based methods are only slightly worse than the supervised ones (GP and relevance vector regression).

Method	**Scientific Podcast**	**Music**	**Radio Drama**	**Average**
Gaussian Process regression	0.52	0.38	0.50	0.47
Relevance Vector regression	0.52	0.44	0.42	0.46
Time-dependent Bayesian CCA	0.42	0.46	0.44	0.44
Bayesian CCA	0.25	0.52	0.44	0.40
Classical CCA	0.35	0.44	0.48	0.42
Random baseline	0.25	0.25	0.25	0.25

6.2 Experiment 2: Inferring Short-Term Attention

The above experiment considered the problem of inferring the level of attention for time-periods lasting roughly one minute, and also matched exactly the experimental setup of our data. For practical application scenarios we might want to infer the level of attention also for shorter time periods, for example to enable auto-pause or audio highlighting, or to more quickly recognize which of several overlapping audio streams the user is attending to.

In this experiment we study how short we can make the time window while still getting an estimate that is better than random chance. We do this by training

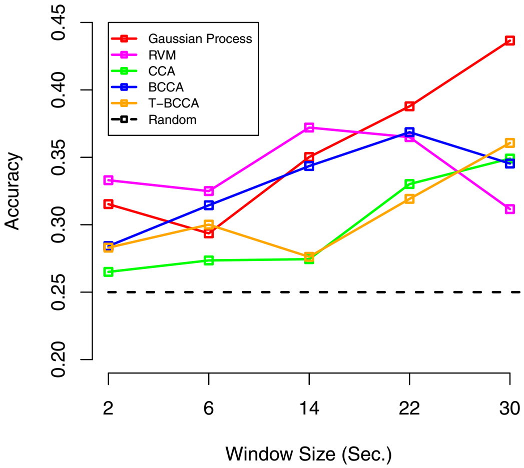


Fig. 4. The classification accuracy of detecting the four levels attention is given for all models in comparison as a function of time window size. For each window size, the scores are averaged over all users and audio types. The unsupervised CCA-based methods are all better than chance, but for most window sizes are slightly beaten by the supervised methods. This figure is best viewed in colors.

the models as above, but making the predictions for all consecutive time windows of certain length[1] instead of the full blocks as we did above. We measure the performance as before, by comparing the true ranking of the M windows with the ranks obtained by ordering the windows based on the correlation score or the regressor output. Again we assume we know how many windows of each attention level we have; this is done for the purpose of measuring only – in practical applications we will always have the full ranking and need not make such assumptions. Figure 4 shows the resulting average accuracy (averaged over both the users and audio types) of the alternative methods as a function of the window size, showing the intuitive trend that inferring the level of attention gets harder when we have less data. Again the supervised predictors are the best, but the unsupervised models also outperform the chance level. For shorter window lengths the Bayesian CCA variants outperform the classical one.

7 Discussion

Visual attention plays a central role in human-computer interaction, and being able to measure the target of the attention with eye-tracking devices has enabled novel types of user interfaces that infer information from the attention [15,20]. Auditory attention, while equally important for the daily life of humans, has been studied much less extensively, not only because auditory interfaces are less

[1] We exclude windows where the ground truth labeling changes during the window.

common but also because no hardware solutions for estimating the level or target of auditory attention exist.

In this work we studied the problem of inferring the level of auditory attention from physiological signals. We compared two alternative approaches for inferring the level: supervised learning with user-independent models, and unsupervised inference based on the assumption that the physiological measurements correlate with the auditory content more strongly when the user is attending to the content. We used state-of-art computational models, Gaussian process [21] and Relevance Vector Machine [23] regression for supervised learning and Bayesian canonical correlation analysis [27] for unsupervised learning, and extended the latter approach to the novel time-dependent BCCA model to better match the underlying time-series nature of the signals. Our experiments demonstrated that both approaches can extract information on the amount of attention paid to three different types of auditory content. The accuracy is not yet sufficient for practical applications, but both approaches outperform chance level with statistical significance, implying that the direction is feasible. The main observation is that the unsupervised methods provide recognition accuracy only slightly worse than that of the supervised models. Even though the time-dependent model was demonstrated to better capture the time-dependencies on artificial generated data, its performance on the real data was only comparable to not modeling the dynamics; the time-dependent model was the best unsupervised variant for long windows, but for short windows the regular Bayesian CCA was better.

One aspect worth noting is that our experiments do not reveal whether the supervised predictors are predicting the level of attention to the auditory content, or merely predicting the attention paid for the visual search tasks. This is because they take as input only the sensory signals and the output labels are equivalent for both tasks. Similar problems are likely to remain for all isolated experiments trying to control the level of auditory attention, and hence for training supervised models guaranteed to address the right aspect one would need to use several alternative techniques for controlling the auditory attention: A supervised model could only be relied to predict the auditory attention itself if it generalizes over all such ways of control. The unsupervised approach, however, does not suffer from the same problem, since we are not learning the parameters to predict the attention but instead are merely estimating the amount of correlation between the sensory signals and the auditory content. This means the approach directly answers to the question of auditory attention, and would not have the flexibility to model alternative explanations for the predicted attention.

For improving the accuracy towards the level required for real-world applications, the most promising direction is to improve the instrumentation and the signal representations. Our main focus was on the machine learning question and the associated computational methods, instead of building a practical attention-detection tool. Replacing the three sensors used in our experiments with sensors more suitable for detecting correlations with the auditory content (for example, a multi-channel EEG additionally recording areas closer to the auditory cortex) should dramatically improve the accuracy, yet the computational methods presented here would remain applicable. Regarding the methods development, a promising direction would be to study methods that allow automatic normalization of the correlation levels with respect to the auditory content. Learning a regressor from the auditory content to the average correlation (independent of the task), would allow normalizing the correlation measures with respect to the content without needing to rely on having measurements from other users.

Acknowledgments. We acknowledge the Academy of Finland (COIN Finnish CoE and decision number 133818) and the UI-ART project for funding. SK and AK were also supported by PASCAL2 European NoE.

References

1. Archambeau, C., Bach, F.: Sparse probabilistic projections. In: Proceedings of NIPS, pp. 73–80 (2009)
2. Bach, F.R., Jordan, M.I.: A probabilistic interpretation of canonical correlation analysis. Tech. Rep. 688, Department of Statistics, University of California, Berkeley (2005)
3. Barber, D., Chiappa, S.: Unified inference for variational Bayesian linear gaussian state-space models. In: Proceedings of NIPS (2006)
4. Bonnel, A.M., Hafter, E.R.: Divided attention between simultaneous auditory and visual signals. Perception & Psychophysics 60(2), 179–190 (1998)
5. Chanel, G., Kronegg, J., Grandjean, D., Pun, T.: Emotion Assessment: Arousal Evaluation Using EEG's and Peripheral Physiological Signals. In: Gunsel, B., Jain, A.K., Tekalp, A.M., Sankur, B. (eds.) MRCS 2006. LNCS, vol. 4105, pp. 530–537. Springer, Heidelberg (2006)
6. Eerola, T., Toiviainen, P.: Mir in matlab: The midi toolbox. In: Proceedings of the International Conference on Music Information Retrieval, ISMIR (2004)
7. Fritz, J., Elhilali, M., David, S., Shamma, S.: Auditory attention–focusing the searchlight on sound. Current Opinions in Neurobiology 17(4), 437–455 (2007)
8. Fujiwara, Y., Miyawaki, Y., Kamitani, Y.: Estimating image bases for visual image reconstruction from human brain activity. In: Procedings of NIPS, pp. 576–584 (2009)
9. Grewal, M.S., Andrews, A.P.: Kalman Filtering: Theory and Practice Using MATLAB. John Wiley and Sons, Inc. (2001)
10. Hardoon, D.R., Szedmak, S., Shawe-Taylor, J.: Canonical correlation analysis: An overview with application to learning methods. Neural Computation 16(12), 2639–2664 (2004)
11. Hillyard, S., Hink, R., Schwent, V., Picton, T.: Electrical signs of selective attention in the human brain. Science 182, 177–180 (1973)
12. Jääskeläinen, I., Ahveninen, P., Bonmassar, G., Dale, A., Ilmoniemi, R., Levanen, S., Lin, F., May, P., Melcher, J., Stufflebeam, S., et al.: Human posterior auditory cortex gates novel sounds to consciousness. Proceedings of National Academy of Science USA 101, 6809–6814 (2004)
13. Kim, J., André, E.: Emotion recognition based on physiological changes in music listening. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(12), 2067–2083 (2008)

14. Klami, A., Kaski, S.: Local dependent components. In: Proceedings of the International Conference on Machine Learning, pp. 425–432. Omnipress (2007)
15. Kozma, L., Klami, A., Kaski, S.: GaZIR: Gaze-based zooming interface for image retrieval. In: Proceedings of the Conference on Multimodal Interfaces (ICMI), pp. 305–312. ACM, New York (2009)
16. Nakai, T., Kato, C., Matsuo, K.: An fMRI study to investigate auditory attention: a model of the cocktail party phenomenon. Magn. Reson. Med Sci. 4(2), 75–82 (2005)
17. Pan, M.K., Chang, G.J.S., Himmetoglu, G.H., Moon, A., Hazelton, T.W., MacLean, K.E., Croft, E.A.: Galvanic skin response-derived bookmarking of an audio stream. In: Proceedings of the Human Factors in Computing Systems (CHI), pp. 1135–1140. ACM, New York (2011)
18. Picard, R.W., Vyzas, E., Healey, J.: Toward machine emotional intelligence: Analysis of affective physiological state. IEEE Trans. Pattern Anal. Mach. Intell. 23(10), 1175–1191 (2001)
19. Pugh, K., Shaywitz, B., Shaywitz, S., Fulbright, R., Byrd, D., Skudlarski, P., Shankweiler, D., Katz, L., Constable, R., Fletcher, J., Lacadie, C., Marchione, K., Gore, J.: Auditory selective attention: An fMRI investigation. Neuroimage 4, 159–173 (1996)
20. Puolamäki, K., Salojärvi, J., Savia, E., Simola, J., Kaski, S.: Combining eye movements and collaborative filtering for proactive information retrieval. In: Proceedings of the International Conference on Research and Development in Information Retrieval (SIGIR), pp. 146–153. ACM, New York (2005)
21. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. MIT Press (2006)
22. Sharp, H., Rogers, Y., Preece, J.: Interaction Design: Beyond Human-Computer Interaction, 2nd edn. John Wiley and Sons (2007)
23. Tipping, M.E.: The relevance vector machine. In: Proceedings of NIPS. MIT Press, Cambridge (2000)
24. Treisman, A.M., Gelade, G.: A feature-integration theory of attention. Cognitive Psychology 12(1), 97–136 (1980)
25. Vertegaal, R., Shell, J.S.: Attentive user interfaces: the surveillance and sousveillance of gaze-aware objects. Social Science Information 47(3), 275–298 (2008)
26. Viinikanoja, J., Klami, A., Kaski, S.: Variational Bayesian Mixture of Robust CCA Models. In: Balcázar, J.L., Bonchi, F., Gionis, A., Sebag, M. (eds.) ECML PKDD 2010, Part III. LNCS, vol. 6323, pp. 370–385. Springer, Heidelberg (2010)
27. Virtanen, S., Klami, A., Kaski, S.: Bayesian CCA via group sparsity. In: Proceedings of the International Conference on Machine Learning (ICML 2011), pp. 457–464. ACM, New York (2011)
28. Wilson, G.F., Russell, C.A.: Real-time assessment of mental workload using psychophysiological measures and artificial neural networks. Human Factors 45(4), 635–643 (2003)