Employing a Binaural Auditory Model to Classify Everyday Sound Events

Simon Ciba[1], Karim Helwani[2], Hagen Wierstorf[2],
Klaus Obermayer[1], Alexander Raake[2], Sascha Spors[2]

[1] *Fachgebiet Neuronale Informationsverarbeitung, TU Berlin, Email: sciba2@hotmail.com*

[2] *Telekom Innovation Laboratories (T-Labs), TU Berlin*

Introduction

Humans benefit considerably from exploiting two ears in everyday listening tasks. It therefore seems to be a promising concept for machine listening approaches to emulate the biological mechanisms of binaural signal processing before applying methods of artificial intelligence. In this work we employ a cross-correlation-based auditory model to automatically perform classification tasks on elementary everyday sound events. We present a heuristic scheme to extract the relevant features from the model's output data. Given a set of training data, a classifier is then constructed using support vector machine (SVM) learning. The proposed method is validated in classification experiments performed on a database of natural sounds. We further discuss its robustness against variation of room acoustics.

Preprocessing by a binaural model

To mimic the outer ears' transfer characteristics and the room acoustic properties of the environment, the incoming sound event is first convolved with a certain binaural room impulse response (BRIR), resulting in two signal streams – one for each ear. Further signal processing is based on the model presented by Lindemann [1] which relies on a *running interaural cross-correlation* process extended by mechanisms of *contralateral inhibition* and *monaural processors*. Prior to the binaural interaction, each ear signal is fed to a peripheral unit consisting of a linear *basilar membrane* filterbank and non-linear innerhaircell transduction implemented as half-wave rectification and low pass filtering at 800 Hz. For each frequency band (index c) the model's output data – denoted here as "activity" $A_c(t,\tau)$ – is given by the (extended) cross-correlation pattern and distributed across two variables: time t and cross-correlation delay τ.

Feature extraction

Given filterbank channel c, all features are extracted from the time-variant peak of activity along the τ-axis. The peak's location is approximated here by the centroid

$$\tau_{0,c}(t) = \frac{\int \tau A_c(t,\tau) d\tau}{\int A_c(t,\tau) d\tau} \qquad (1)$$

with the integral extending across the interval of valid τ-values, and is conceptually considered as a criterion for lateralization [1]. Temporal fluctuations of $\tau_{0,c}$ may also be related to the perception of source width [2]. As more relevant for event classification, however, may be regarded the peak's height as defined by

$$\tilde{A}_c(t) = A_c(t, \tau_{0,c}(t)), \qquad (2)$$

which is strongly associated with the product of the envelopes of both ear signals, smoothed by weighted integration throughout the correlation window. It therefore to some extent represents a distribution of energy across time and frequency bands. Intuitively, the peak's width may contain information about the amount of randomness inherent in the source signal, and hence we calculate the (squared) spread of activity around the centroid as

$$\sigma_c^2(t) = \frac{\int A_c(t,\tau) \cdot (\tau - \tau_{0,c}(t))^2 d\tau}{\int A_c(t,\tau) d\tau}. \qquad (3)$$

To reduce the complexity of the subsequent learning problem, similar as in [3] we perform feature integration via computation of mean and variance from all three predefined time series and, additionally, from their first derivatives. This brings the time axis down to 4 points, in total leading to a feature space of $3 \times 4 \times N_C$ dimensions, where N_C denotes the number of filterbank channels.

Learning of the classifier

A binary classifier is learned by application of the C-SVM ansatz proposed in [4] which takes into account different class sizes in the training dataset by weighting the regularization parameter, usually designated by the letter C, accordingly. From preliminary experiments [5] we know that a polynomial kernel function

$$k(\mathbf{x}^{(\alpha)}, \mathbf{x}^{(\beta)}) = \left[\left(\mathbf{x}^{(\alpha)} \right)^T \mathbf{x}^{(\beta)} + 1 \right]^d \qquad (4)$$

with $d \geq 1$ being the degree of the polynomial and $\mathbf{x}^{(\alpha)}$, $\mathbf{x}^{(\beta)}$ being two points from feature space, provides a suitable trade-off between complexity and generalizability of the learning process. To select appropriate values of the hyperparameters C and d, we browse the grid spanned by $2^{-6} \leq C \leq 2^{10}$ and $1 \leq d \leq 5$ in advance, and pick the values that lead to the best outcome of a stratified 5-fold-cross-validation based on the entire training data.

Experiments

Data corpora

To evaluate our approach, we resort to a sound samples database (see Table 1) that has been presented in [6] and is taxonomically categorized by the sound generating interaction of materials and their aggregate states. The samples are monophonic recordings with a sampling rate of 11025 Hz, and either encompass a single acoustic event or have a repetitive or continuous temporal structure. In the latter two cases signals are cut to a length of 4 s. For the BRIRs we utilize the recordings from an anechoic chamber ("AC"), a control room of a recording studio ("ST"), a conference room ("CR") and a meeting room ("MR"). While the first one exhibits only weak reflections such that its BRIR can be interpreted as a head-related impulse response (HRIR), the reverberation times T_{60} of the last three rooms are 0.2 s, 0.4 s and 0.6 s, respectively.

Table 1: Database of elementary everyday sounds.

category	class	no.	temporal structure
solid	impact	92	single
	rolling	116	continuous
	deformation	55	single
	friction	70	continuous
gas	wind	46	continuous
	whoosh	50	single
	explosion	81	single
liquid	drip	39	repetitive/continuous
	flow	51	continuous
	pour	80	continuous

Methodology

From the 10 sound classes listed in Table 1 we construct 45 binary classification problems in a *one-against-one* manner by combining each class with each of the remaining classes individually. Besides, we gather the remaining classes in a rest-class and this way yield further 10 problems in a *one-against-rest* manner. For each problem we run a stratified 10 × 10-cross-validation yielding an average balanced prediction accuracy value. Since we are interested in a general statement, we calculate the mean and standard deviation of the results across problems. To assess robustness, the described validation process is repeated for each of the four BRIRs, which then is applied to the data in the prediction step. For this approach we further distinguish three scenarios according to different qualities of learning: in the first case ("A") the room employed for prediction is also underlying each instance of the training data. In the second ("B"), learning takes place on a random mixture of all four rooms and in the third case ("C"), the mixture lacks the targeted room.

Implementation

We use a MATLAB implementation of Lindemann's model as part of the Auditory Modelling Toolbox (AMT, v0.02) [7]. The quadratic optimization problem in the C-SVM ansatz is solved by the LIBSVM software package [8]. Selected parameter values and further technical details are documented in [5].

Results

Results are shown in Table 2. Under free-field conditions mean accuracy lies at about 95 % in case of one-against-one tasks and at 90 % for one-against-rest tasks. For the three remaining rooms the results are comparable if the target room has been subjected to the entire training data. Variation of room acoustics leads to a decrease in mean accuracy of up to 5.3 % and 6.9 %, respectively, if the targeted room is included in learning. If the room is excluded, results become substantially worse.

Table 2: Mean μ and standard deviation σ of the validation results across problems.

			AC	ST	CR	MR
A	1 vs. 1	μ	95.41	95.54	95.27	94.14
		σ	4.62	4.48	4.94	5.46
	1 vs. Rest	μ	90.00	90.37	89.95	87.65
		σ	4.99	5.24	5.55	5.59
B	1 vs. 1	μ	92.42	90.24	92.36	91.30
		σ	6.79	7.03	6.50	6.68
	1 vs. Rest	μ	86.27	83.55	85.09	83.56
		σ	6.50	7.21	7.17	7.10
C	1 vs. 1	μ	70.85	67.66	88.75	84.66
		σ	12.52	13.22	8.17	10.57
	1 vs. Rest	μ	64.96	63.11	79.17	72.50
		σ	9.45	10.17	9.95	7.12

Conclusion

Preprocessing audio data by a binaural model can be understood as an important step into emulating the human perception of an auditory event. Practical experiments gave a proof of concept and have shown that the presented attempt to audio classification yields reasonable results, if the room acoustics underlying prediction are known at learning time. It may be suspected that the latter restriction will become less important as variety of rooms in the training data increases. Assuming the existence of a sufficiently large database of BRIRs, future research may include "binaural" approaches to the classification of room acoustic environments or – as a more general case – to the prediction of spatial sound attributes.

Acknowledgment

This work was funded by the Innovation Center Human-Centric Communication, Technische Universität Berlin.

References

[1] Lindemann, W.: Extension of a binaural cross-correlation model by contralateral inhibition. I. Simulation of lateralization for stationary signals. Journal of the Acoustical Society of America 80 (1986), 1608-1622.

[2] Hess, W.: Time-variant binaural-activity characteristics as indicator of auditory spatial attributes. Dissertation, Institute of Communication Acoustics, Ruhr Universität, Bochum, 2006.

[3] Adiloğlu, K., Anniés, R., Purwins, H., Obermayer, K.: Closing the Loop of Sound Evaluation and Design (CLOSED), Deliverable 5.1: Representations and Predictors for Everyday Sounds. Neural Information Processing Group, Technische Universität Berlin, 2008.

[4] Huang, Y.-M., Du, S.-X.: Weighted Support Vector Machine for Classification with uneven Training Class Sizes. Proceedings of the Fourth International Conference on Machine Learning and Cybernetics (2005), 4365-4369.

[5] Ciba, S.: Automatic Classification of Everyday Sound Events after Preprocessing by a Binaural Auditory Model. Studienarbeit, Technische Universität Berlin, 2011.

[6] Adiloğlu, K., Anniés, R., Purwins, H., Obermayer, K.: Closing the Loop of Sound Evaluation and Design (CLOSED), Deliverable 5.2: Visualisation and Measurement Assisted Design. Neural Information Processing Group, Technische Universität Berlin, 2009.

[7] Søndergaard, P.L., Culling, J.F., Dau, T., Le Goff, N., Jepsen, M.L., Majdak, P., Wierstorf, H.: Towards a binaural modelling toolbox. Proceedings of the Forum Acusticum 2011 (2011).

[8] Chang, C.-C., Lin, C.-J.: LIBSVM: a library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm (2001).