UNIVERSITÄT ZU LÜBECK
INSTITUTE FOR SIGNAL PROCESSING

Learning Compact Structural Representations for Audio Events Using Regressor Banks

Huy Phan, Marco Maass, Lars Hertel, Radoslaw Mazur, Ian McLoughlin, *and* Alfred Mertins

1. Introduction

We introduce a novel descriptor learned by a bank of random regression forests for audio event representation. The descriptor offers different advantages:

- **Temporal encoding**: the temporal structure of an audio event category is modeled by a class-specific regression forest in the bank.
- **Shared feature encoding**: the responses of the regressor bank on a target event quantify how it aligns to the structures of different event classes.
- **Compact**: the number of entries equals the number of event categories.
- **Discriminative**: state-of-the-art performance even with linear classifiers.

2. Random Regression Forest for Temporal Encoding

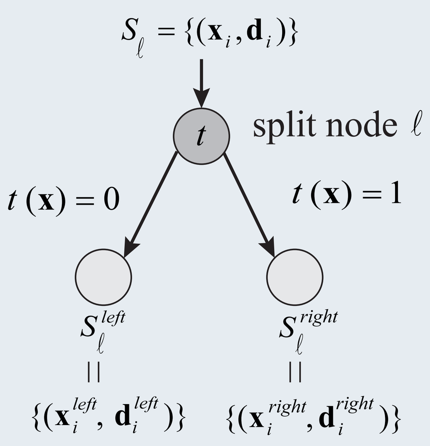
Training

- Training audio events are decomposed into a set of audio segments $\mathcal{S}: \{s_i = [\mathbf{x}_i, \mathbf{d}_i]; i = 1 \dots |\mathcal{S}|\}$, where
  - $\mathbf{x}_i \in \mathbb{R}^M$: feature vector
  - $\mathbf{d}_i = [d_i^+, d_i^-] \in \mathbb{R}_+^2$: distance vector to event onset and offset
- Tree construction [4]
  - Binary test at split nodes:

$$t_{r,\tau}(\mathbf{x}) = \begin{cases} 1, & \text{if } \mathbf{x}^r > \tau \\ 0, & \text{otherwise.} \end{cases}$$

  - The optimal test is chosen by:

$$t^*_{r,\tau} = \operatorname{argmin}_{t_{r,\tau}} \left(\sum_i \|\mathbf{d}_i^{\text{left}} - \bar{\mathbf{d}}^{\text{left}}\|_2^2 + \sum_i \|\mathbf{d}_i^{\text{right}} - \bar{\mathbf{d}}^{\text{right}}\|_2^2 \right).$$



- Onset and offset distances at a leaf are modeled as Gaussians $\mathcal{N}^+(\bar{d}^+, \Sigma^+)$ and $\mathcal{N}^-(\bar{d}^-, \Sigma^-)$ where $\bar{d}$ and Σ, respectively, denote the mean and variance.

Testing

- Event onset and offset estimations by a tree given a test audio segment $\mathbf{x}_{n'}$ at the time index n':

$$p^+(n|\mathbf{x}_{n'}, \bar{d}^+, \Sigma^+) = \mathcal{N}^+(n; n' - \bar{d}^+, \Sigma^+), \quad (1)$$
$$p^-(n|\mathbf{x}_{n'}, \bar{d}^-, \Sigma^-) = \mathcal{N}^-(n; n' + \bar{d}^-, \Sigma^-). \quad (2)$$

- Event onset and offset estimations by the forest of T trees:

$$p^+(n|\mathbf{x}_{n'}) = \frac{1}{T}\sum_{t=1}^{T} p^+(n|\mathbf{x}_{n'}, \bar{d}_t^+, \Sigma_t^+), \quad (3)$$
$$p^-(n|\mathbf{x}_{n'}) = \frac{1}{T}\sum_{t=1}^{T} p^-(n|\mathbf{x}_{n'}, \bar{d}_t^-, \Sigma_t^-). \quad (4)$$

3. Bank-of-regressors (BoR) descriptor

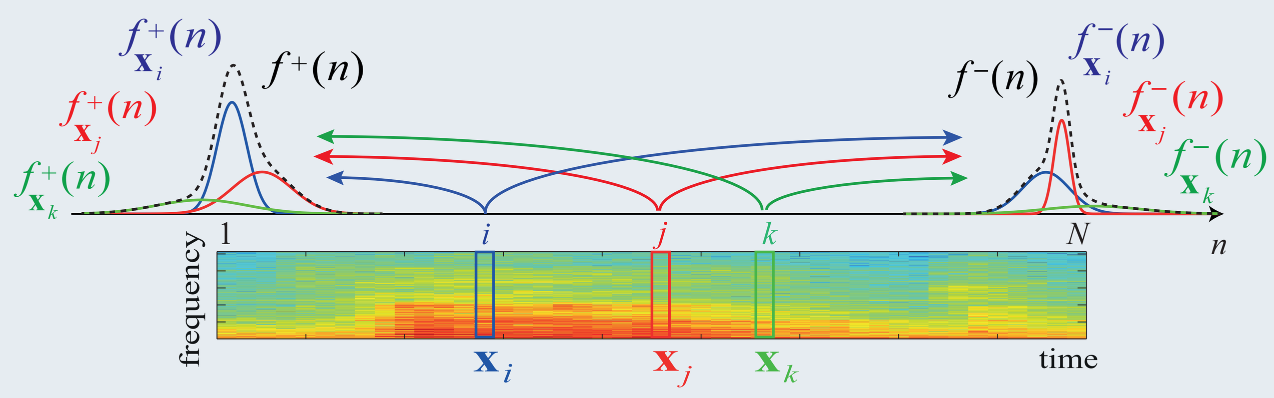


Figure 1. Temporal coding for an audio event

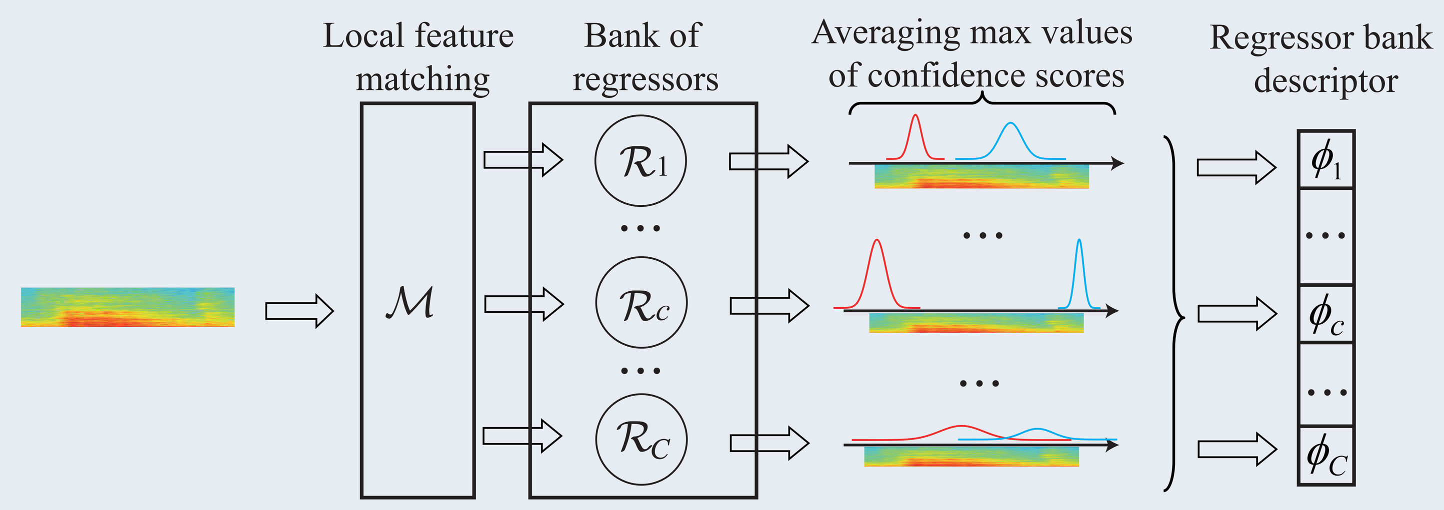


Figure 2. Extraction of BoR descriptor

- A sequence of audio segments $(\mathbf{x}_n; n = 1 \dots N)$ of a target event is transformed into a compact BoR descriptor $\phi = [\phi_1, \dots, \phi_C]^T \in \mathbb{R}_+^C$, where

$$\phi_c = \frac{1}{2}\left(\max_n \left(f_c^+(n)\right) + \max_n \left(f_c^-(n)\right)\right), \quad (5)$$

$$f_c^+(n) = \sum_{i=1}^{N} p^+(n, c|\mathbf{x}_i) = \sum_{i=1}^{N} P(c|\mathbf{x})p^+(n|\mathbf{x}, c), \quad (6)$$
$$f_c^-(n) = \sum_{i=1}^{N} p^-(n, c|\mathbf{x}_i) = \sum_{i=1}^{N} P(c|\mathbf{x})p^-(n|\mathbf{x}, c). \quad (7)$$

  - $c \in \{1, \dots, C\}$ where C is the number of target event categories
  - $p^+(n|\mathbf{x}, c)$ and $p^-(n|\mathbf{x}, c)$ given in (3) and (4), respectively
  - $P(c|\mathbf{x})$ is the probability that segment $\mathbf{x}$ matches to event class c, which is modeled by the random forest classifier $\mathcal{M}$
- Fusion of structural and non-structural descriptors
  - Non-structural features $\varphi = [\varphi_1, \dots, \varphi_C]^T \in \mathbb{R}_+^C$, where

$$\varphi_c = \frac{1}{N}\sum_{n=1}^{N} P(c|\mathbf{x}_n). \quad (8)$$

  - Fusion of two descriptors with extended Gaussian kernel:

$$K(e_i, e_j) = \exp\left(-\sum_{k\in\{\phi,\varphi\}} \frac{1}{A^k} D(e_i^k, e_j^k)\right), \quad (9)$$

where $D(e_i^k, e_j^k)$ is χ^2 distance between audio events e_i and e_j on k-th channel and A^k is mean of D in training data.

4. Experimental results

Setup

- **Databases**: ITC-Irst, UPC-TALP, Freiburg-106, NAR.
- **Low-level features**: 16 log-frequency filter bank parameters + Δ + $\Delta\Delta$, zero-crossing rate, short time energy, four sub-band energies, spectral flux, spectral centroid, and spectral bandwidth.
- **Our classifiers**: linear SVM with BoR descriptors (**BoR-linear**), χ^2-kernel SVM with BoR descriptors (**BoR-χ^2**), and SVM with feature fusion (**BoR+**).
- **Baselines**: Bag-of-words (**BoW**), pyramid BoW (**PBoW**), and **max voting**.

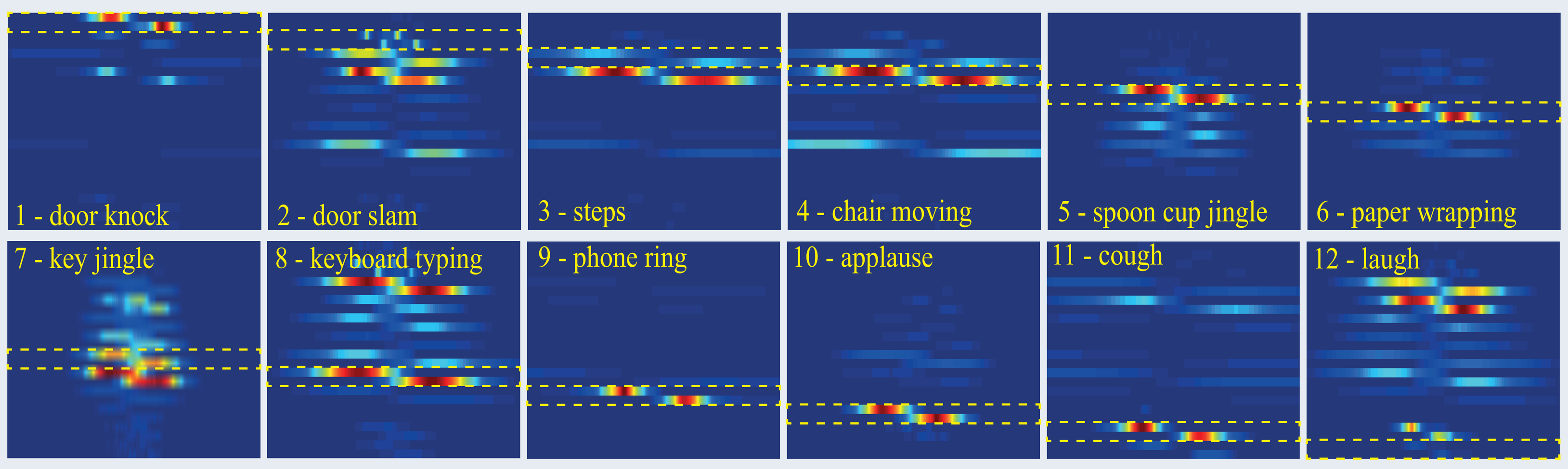


Figure 3. Responses of regressor bank on audio events of different classes.

Experimental Results

Table 1. Overall f-score (%) with the segment size of 50 ms.

Dataset	BoW	PBoW	Max voting	Best reported	Our systems		
					BoR-linear	BoR-χ^2	BoR+
ITC-Irst	97.3	96.6	95.9	97.3 [5]	**97.9**	**97.9**	**99.3**
UPC-TALP	96.6	96.5	94.5	87.6 [2]	95.8	**96.7**	**96.8**
Freiburg-106	96.6	96.8	92.3	**98.9** [3]	97.2	97.8	98.1
NAR	94.8	96.4	92.6	97.0 [1]	96.8	**97.6**	**97.6**

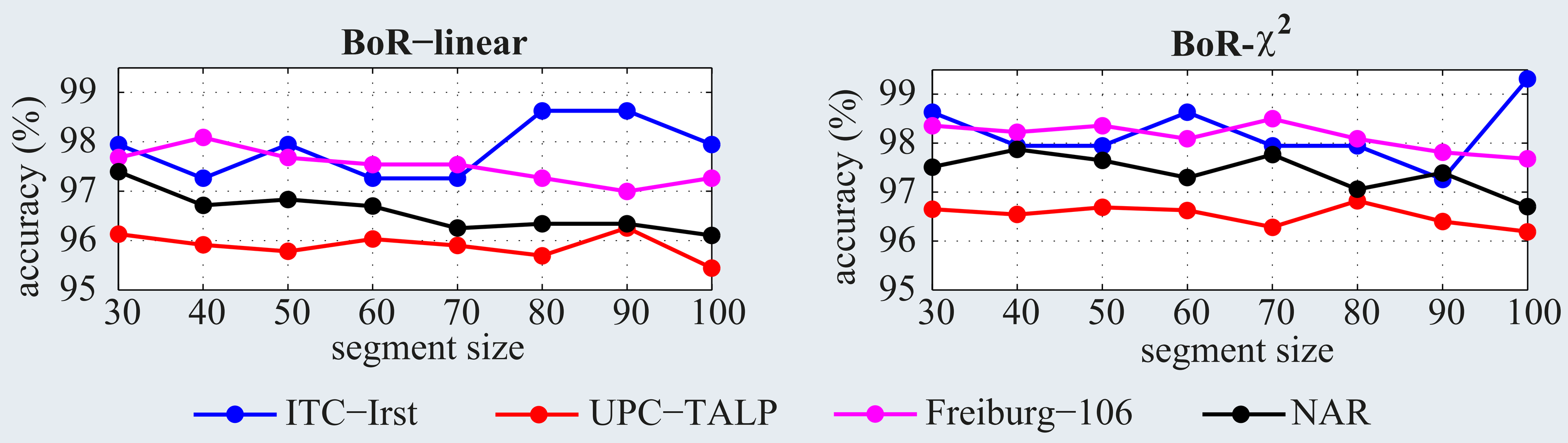


Figure 4. Classification accuracy as a function of audio segment size.

References

[1] J. Maxime, X. Alameda-Pineda, L. Girin, and R. Horaud. Sound representation and classification benchmark for domestic robots. In *Proc. ICRA*, pages 6285–6292, 2014.
[2] C. Nadeu, R. Chakraborty, and M. Wolf. Model-based processing for acoustic scene analysis. In *Proc. EUSIPCO*, pages 2370–2374, 2014.
[3] H. Phan, L. Hertel, M. Maass, and A. Mertins. Audio phrases for audio event recognition. In *Proc. EUSIPCO*, pages 2546–2550, 2015.
[4] H. Phan, M. Maaß, R. Mazur, and A. Mertins. Random regression forests for acoustic event detection and classification. *TASLP*, 23(1):20–31, 2015.
[5] H. Phan and A. Mertins. Exploring superframe co-occurrence for acoustic event recognition. In *Proc. EUSIPCO*, pages 631–635, 2014.