REGULARIZED SPARSE KERNEL SLOW FEATURE ANALYSIS

Wendelin Böhmer[1,2], Steffen Grünewälder[1,3], Hannes Nickisch[1,4] and Klaus Obermayer[1,2]

Bernstein Focus: Neurotechnology Berlin

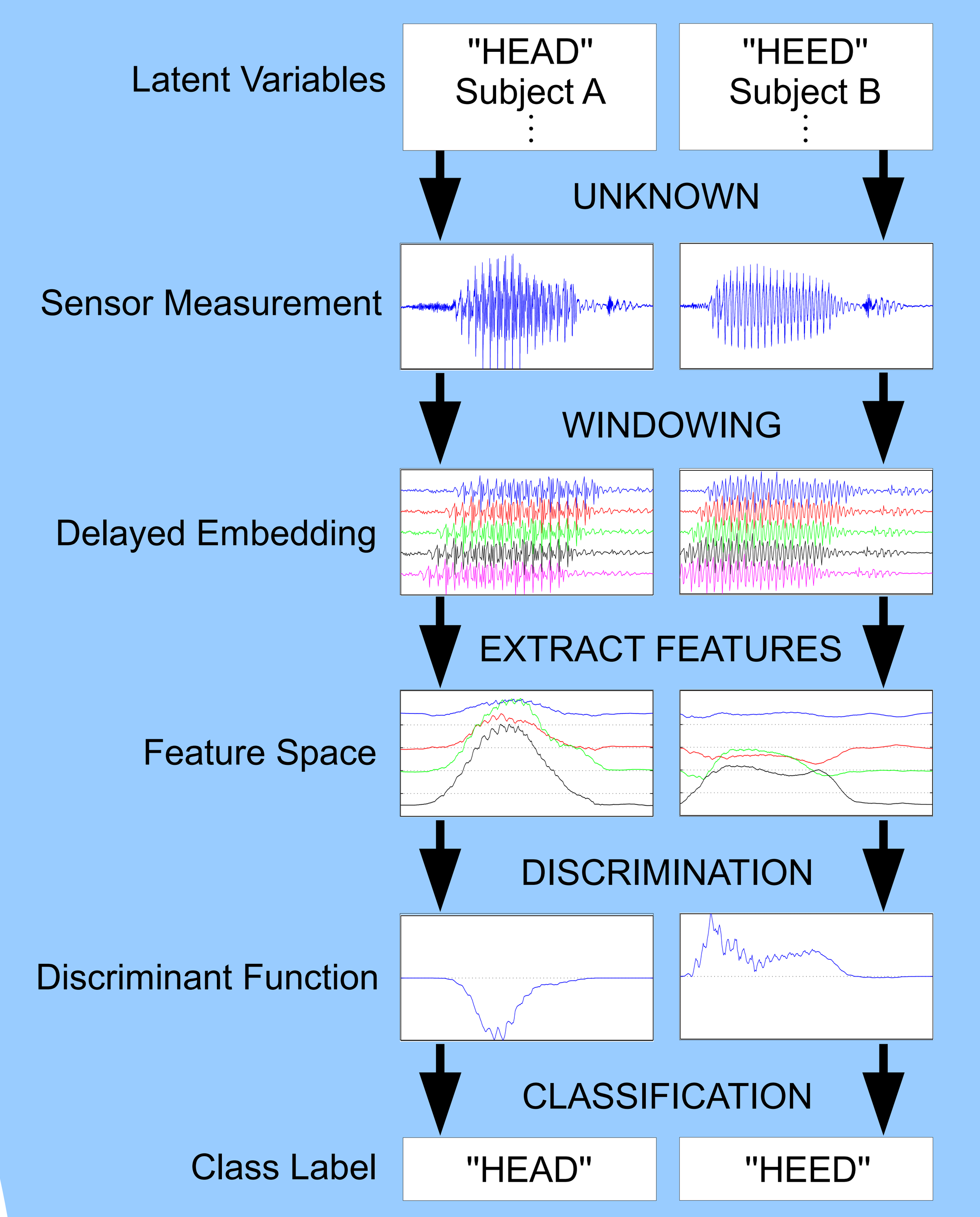
(1) Neural Information Processing Group, Technische Universität Berlin, Germany
(2) Bernstein Focus: Neurotechnologie, Bernstein Center Berlin, Germany
(3) Centre for Computational Statistics and Machine Learning, University College London, United Kingdom
(4) Philips Research Laboratories, Hamburg, Germany

ABSTRACT

This paper develops a kernelized slow feature analysis (SFA) algorithm. SFA is an unsupervised learning method to extract features which encode latent variables from time series. Generative relationships are usually complex, and current algorithms are either not powerful enough or tend to over-fit. We make use of the kernel trick in combination with sparsification to provide a powerful function class for large data sets. Sparsity is achieved by a novel matching pursuit approach that can be applied to other tasks as well. For small but complex data sets, however, the kernel SFA approach leads to over-fitting and numerical instabilities. To enforce a stable solution, we introduce regularization to the SFA objective. Feature extraction is demonstrated on a vowel classification task.

UNSUPERVISED NON-LINEAR FEATURE EXTRACTION

- **Classification** or **regression** w.r.t. **latent variables**
  - Latent variables **non-linearly embedded** in data
  - Discriminant/regression function **non-linear** in the space of latent variables
- Utilize knowledge from **unlabelled data**
  - **Unsupervised** construction of features from data
  - Functional basis on **manifold** of latent variables
- Non-linear **PCA** does not encode latent variables
- Non-linear **SFA** does



OBJECTIVE

SLOW FEATURE ANALYSIS

$$\min \sum_i E_t\left[\dot{y}_i^2(x_t)\right]$$

$$\text{s.t.} \quad \forall i: \; E_t\left[y_i(x_t)\right] = 0$$

$$\forall i,j: \; E_t\left[y_i(x_t)\, y_j(x_t)\right] = \delta_{ij}$$

$y_i(x)$ can be from an arbitrary function class

Given an **infinite time series** and **sufficient function class**, SFA features span a **Fourier basis** in the **space of the latent variables** [I].

FUNCTION CLASS

REPRODUCING KERNEL HILBERTSPACES

$$\min \; \tfrac{1}{n-1} \mathbf{tr}\left(A^T K D D^T K^T A\right)$$

$$\text{s.t.} \quad \tfrac{1}{n} A^T K \mathbf{1} = \mathbf{0}$$

$$\tfrac{1}{n} A^T K K^T A = I$$

$$D = \begin{bmatrix} -1 & 0 & \cdots & 0 \\ 1 & -1 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -1 \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

$$y_i(x) = \sum_t A_{ti}\, \kappa(x, x_t) - c_i$$

Only **small training-sets** feasible due to complexity. For any **finite training-set**, kernel SFA exhibits **over-fitting** and **numerical instability** when applied on a test-set [III].

Regularized Sparse Kernel

RSK-SFA

Slow Feature Analysis

REGULARIZATION

PENALIZE COMPLEX FUNCTIONS BY PENALIZING HILBERT NORM

$$\min \sum_i \left(E_t\left[\dot{y}_i^2(x_t)\right] + \lambda \|y_i\|_H^2\right) \equiv$$

$$\min \; \tfrac{1}{n-1} \mathbf{tr}\left(A^T K D D^T K^T A\right) + \lambda\, \mathbf{tr}\left(A^T \bar{K} A\right)$$

$$\bar{K}_{ij} = \kappa(z_i, z_j)$$

Optimal constant λ can be **very small**.
Adapts better to the objective.
No computational overhead.
No speed-up.

SPARSENESS

PREVENT COMPLEX FUNCTIONS BY RESTRICTION TO SUBSPACE

$$y_i(x) \in \mathbf{span}\left(\{\kappa(x, z_j)\}_{j=1}^m\right)$$

$$\{z_j\}_{j=1}^m \subset \{x_t\}_{t=1}^n, \qquad m \ll n$$

$$K_{jt} = \kappa(z_j, x_t)$$

Removes the computational **bottle-neck**.
Regularizes the solution **indirectly**.
Depends strongly on **sparse subset selection**.

SPARSE SUBSET SELECTION

- $\{z_j\}_{j=1}^m$ should span the maximal possible subspace

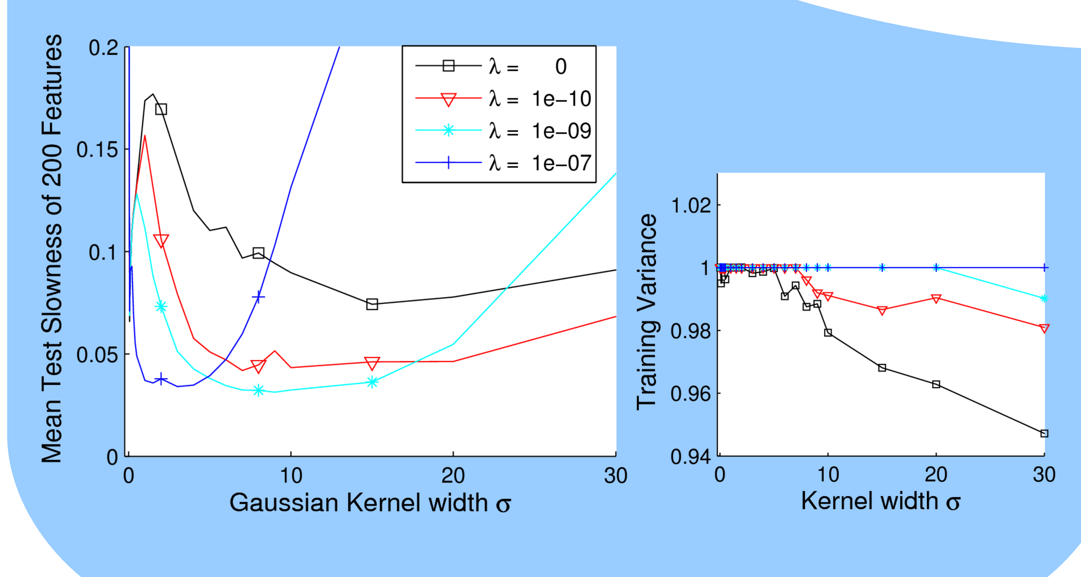
$$\epsilon_t^i := \min_\alpha \left\| \kappa(\cdot, x_t) - \sum_{j=1}^m \alpha_j \kappa(\cdot, x_{i_j}) \right\|^2$$

- Equivalent to the sparse kernel PCA problem
  - *Matchig Pursuit for Sparse Kernel PCA* [V] $O(n^2 m)$
  - *Online Maximization of the Affine Hull* [IV] $O(n m^2)$
- *Matching Pursuit of Online MAH* $O(n m^2)$

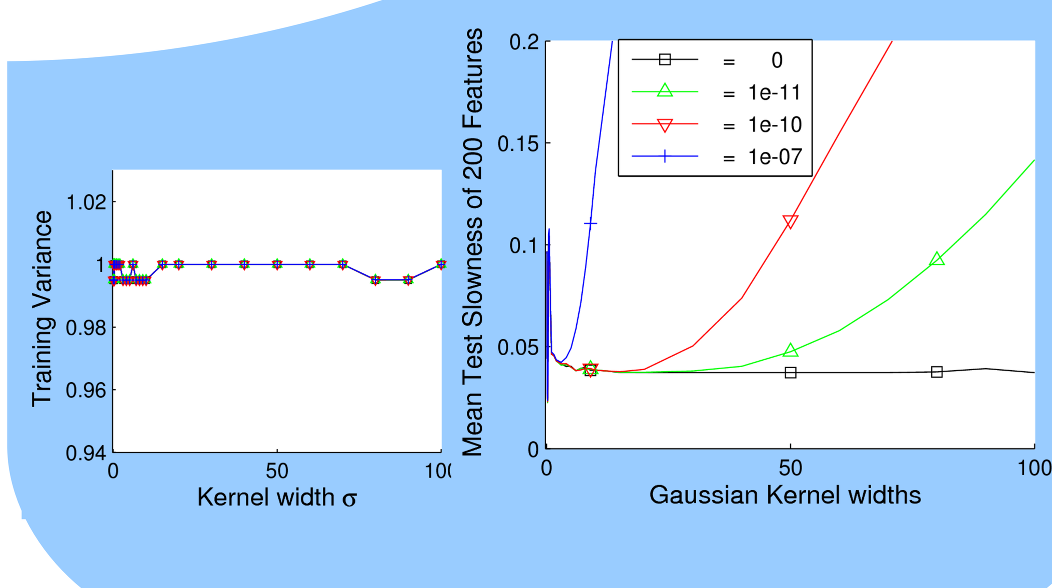
$$\boldsymbol{i}_{j+1} := \mathit{argmin}_t \left\| [\epsilon_1^{i \cup t}, \ldots, \epsilon_n^{i \cup t}] \right\|_\infty \approx \mathit{argmax}_t \, \epsilon_t^i$$

$$\epsilon_t^{i \cup j} = \epsilon_t^i - \frac{1}{\epsilon_j^i}\left(K_{tj} - \boldsymbol{K}_{ti}(\boldsymbol{K}_{ii})^{-1} \boldsymbol{K}_{ij}\right)^2$$

RESULTS

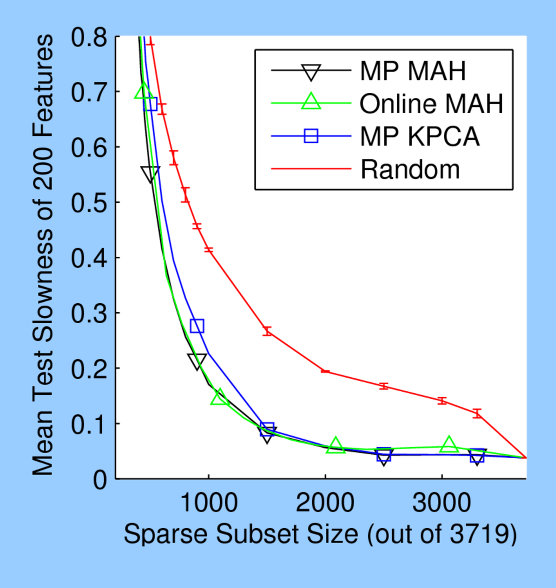


RESULTS

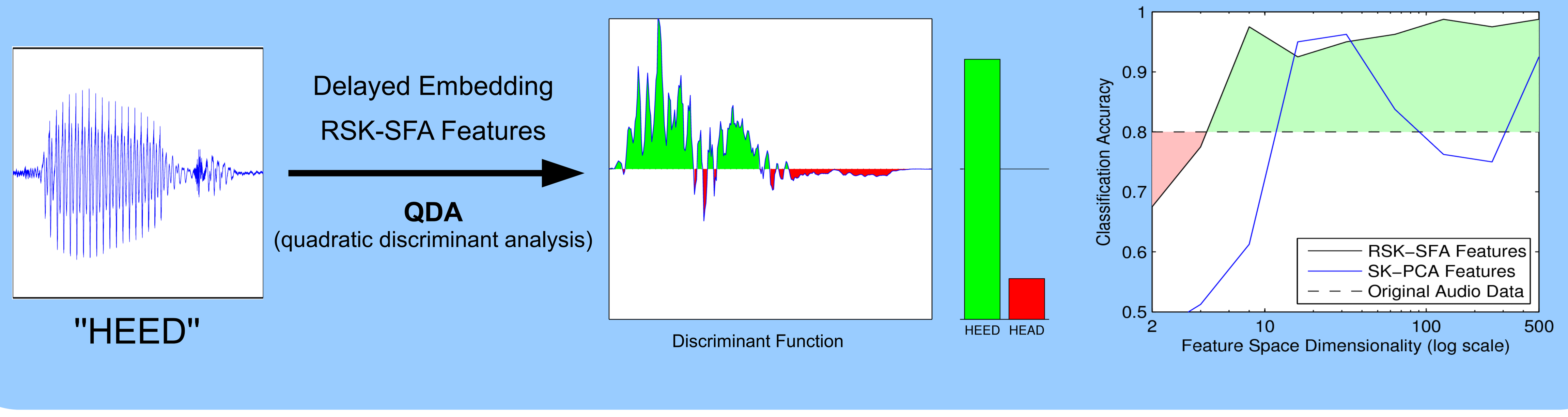


RESULTS

- All algorithms outperform random selection
- MP MAH exhibits **same performance**
- MP MAH returns **ordered subset**
- **Regularization** by **sparseness** becomes much easier



FEATURE VALIDATION: VOWEL CLASSIFICATION



REFERENCES

I.Wiskott: *Slow feature analysis: a theoretical analysis of optimal free responses.* Neural Computation 15(4):2147-2177, (2002)
II.Bray and Martinez: *Kernel based extraction of slow features: complex cells learn disparity and translation invariance from natural images.* NIPS 15:253-260, (2002)
III.Fukumizu, Bach and Gretton: *Statistical consistency of kernel canonical correlation analysis.* Journal of Machine Learning Research 8:361-383, (2007)
IV.Csato and Opper: *Sparse on-line gaussian processes.* Neural Computation 14(3):641-668, (2002)
V.Hussain and Shawe-Taylor: *Theory of matching pursuit.* NIPS 21:721-728, (2008)

contact: Wendelin Böhmer `<wendelin@cs.tu-berlin.de>`

This work has been supported by the IGP on Human-Centric Communication at TU Berlin and the German Federal Ministry of Education and Research (grant 01GQ0850)