# Analysing audio files by the example of birdsong recognition

## Strukturiertes Promotionsprogramm Data Science

Sven Heuer

Philipps-University Marburg

June 2nd, 2021

# Motivation

Project: Nature 4.0 | Sensing Biodiversity

45 microphones recording ∼10h/day, leading to around 1TB of audio data per week.

# Challenges

- Denoise data
- Compress data
- Detect relevant parts in audio files
- Classify birds
- Deploy algorithms

# Why not Wavelets?

Dilation and Translation operator:

$$D_a f(t) = |a|^{-1/2} f(t/a), \; T_x f(t) = f(t - x)$$

Wavelet transform:

$$W_g f(x, a) = \langle f, T_x D_a g \rangle$$

[1]https://in.mathworks.com/help/wavelet/ug/wavelet-analysis-of-physiologic-signals.html

# Why not Wavelets?

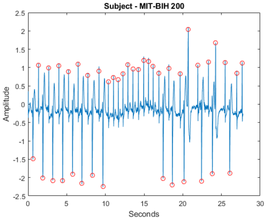
Dilation and Translation operator:

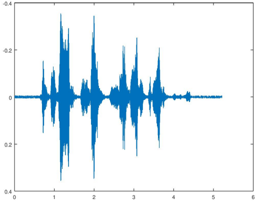
$$D_a f(t) = |a|^{-1/2} f(t/a), \;\; T_x f(t) = f(t-x)$$

Wavelet transform:

$$W_g f(x,a) = \langle f, T_x D_a g \rangle$$

Different signal types[1]:




[1] https://in.mathworks.com/help/wavelet/ug/wavelet-analysis-of-physiologic-signals.html

# The short time Fourier transform

Modulation operator:

$$M_\omega f(t) = e^{2\pi i\omega t} f(t)$$

STFT:

$$V_g f(x,\omega) = \langle f, M_\omega T_x g \rangle$$

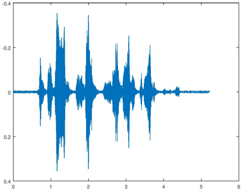
# The short time Fourier transform

Modulation operator:

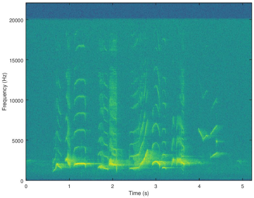
$$M_\omega f(t) = e^{2\pi i \omega t} f(t)$$

STFT:

$$V_g f(x,\omega) = \langle f, M_\omega T_x g \rangle$$




$\longrightarrow$



# Gabor Frames

## Definition

Consider the grid $\Lambda = \alpha\mathbb{Z}^d \times \beta\mathbb{Z}^d$ and a window function $g \in L_2(\mathbb{R}^d)$. The set

$$\{g_\lambda = (T_{\alpha k} M_{\beta n} g) \mid \lambda = (\alpha k, \beta n) \in \Lambda\}$$

is called a *Gabor frame*, if $0 < A \leq B < \infty$ exist, such that

$$A \|f\|^2 \leq \sum_{\lambda \in \Lambda} |\langle f, g_\lambda \rangle|^2 \leq B \|f\|^2$$

holds for all $f \in L_2(\mathbb{R}^d)$.

Rule of thumb: If $\alpha$ and $\beta$ are „small enough“, the Gabor system is a frame.

# Frame Reconstruction

## Lemma

*Let* $\{g_\lambda \mid \lambda \in \Lambda\}$ *be a Gabor frame. Then there exists a dual window* $\tilde{g}$*, such that* $\{\tilde{g}_\lambda \mid \lambda \in \Lambda\}$ *forms a frame and every* $f \in L_2(\mathbb{R}^d)$ *can be written as*

$$f = \sum_{\lambda \in \Lambda} \langle f, \tilde{g}_\lambda \rangle g_\lambda$$

*or*

$$f = \sum_{\lambda \in \Lambda} \langle f, g_\lambda \rangle \tilde{g}_\lambda .$$

# Modulation Spaces

## Definition

For $0 \neq g \in \mathcal{S}(\mathbb{R})$. Modulation space:

$$\mathcal{M}_p(\mathbb{R}) = \left\{ f : \mathbb{R} \to \mathbb{C} \;\middle|\; V_g f \in L_p(\mathbb{R}^2) \right\}.$$

Modulation norm:

$$\|f\|_{\mathcal{M}_p(\mathbb{R})} = \left( \int_{\mathbb{R}} \int_{\mathbb{R}} |V_g f(x, \omega)|^p \, \mathrm{d}x \mathrm{d}\omega \right)^{1/p}.$$

Remark: Using this „template“ with wavelets, we get Besov spaces.

## Compression

Preconditions:

- $p \in (0,2)$
- $\Lambda$ grid, so that Gabor atoms form a frame
- $g \in \mathcal{M}_p(\mathbb{R})$, $\tilde{g}$ dual window
- $f \in \mathcal{M}_p(\mathbb{R})$

### Satz

*For* $\mu > 0$, *let* $I_\mu = \{\lambda \in \Lambda \mid |\langle f, \tilde{g}_\lambda \rangle| \geq \mu\}$ *and*

$$f_\mu = \sum_{\lambda \in I_\mu} \langle f, \tilde{g}_\lambda \rangle g_\lambda.$$

*Then (with* $N = \# I_\mu$*):*

$$\|f - f_\mu\|_{L_2}^2 \leq C \, \|f\|_{\mathcal{M}_p(\mathbb{R})}^2 \, N^{1-2/p}.$$

With a rudimentary implementation: Compression with factor 7.

# B-Splines

- Necessary condition: $g \in \mathcal{M}_p(\mathbb{R})$
- Gaussian window is even in $\mathcal{S}(\mathbb{R})$, but expensive.

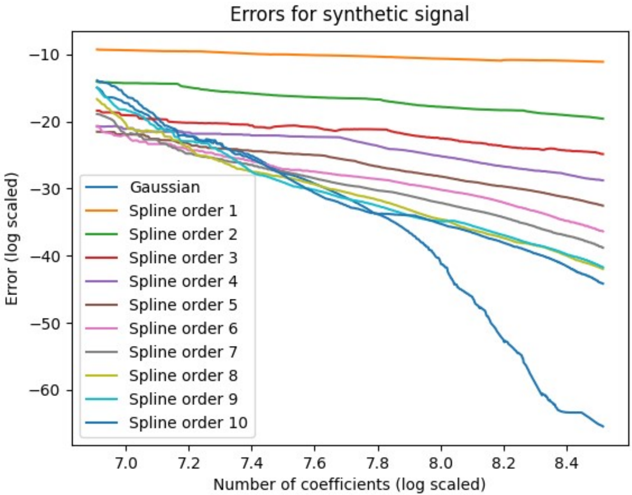
## Satz

*Let $g$ be a B-Spline of order $k$. Then*

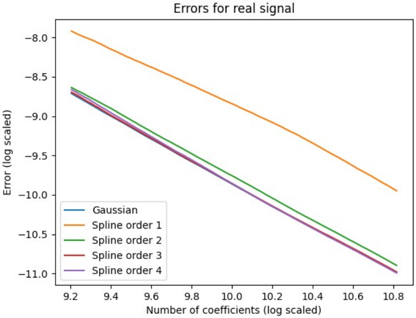
$$g \in \mathcal{M}_p(\mathbb{R})$$

*for all $p > 1/k$.*

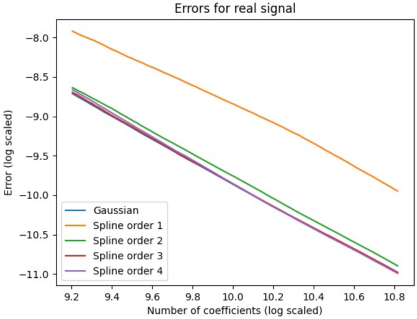
# Simulations - synthetic signal



# Simulations - real signal



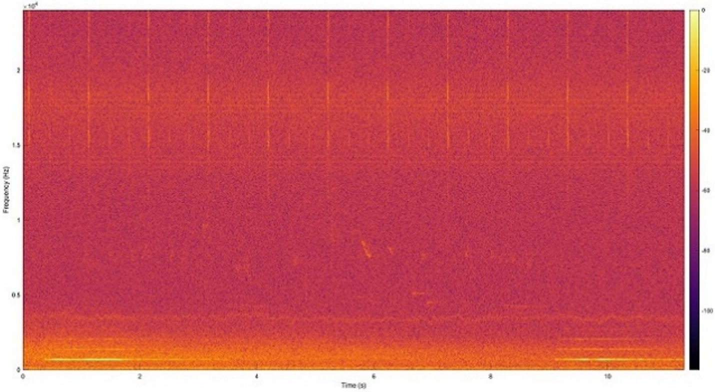
# Simulations - real signal




| Window | $p$ (simulation) |
|---|---|
| Spline order 1 | 0.8844 |
| Spline order 2 | 0.8322 |
| Spline order 3 | 0.8317 |
| Spline order 4 | 0.8207 |
| Gaussian | 0.8342 |

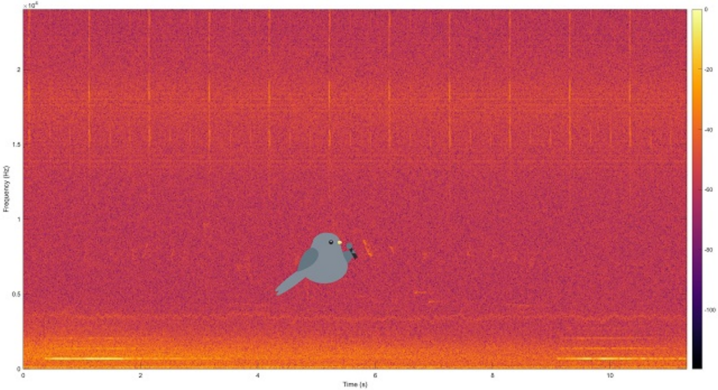
# Usability

Recording from the forest (trains, train whistle, bird):



# Usability

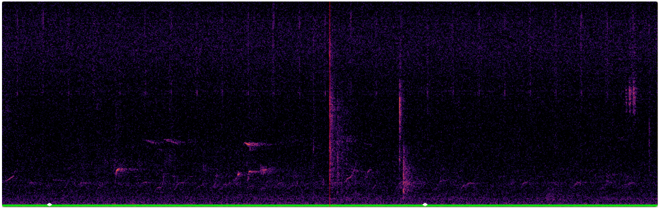
Recording from the forest (trains, train whistle, bird)[2]:




[2]https://pixabay.com/illustrations/sing-bird-songbird-microphone-1322180/

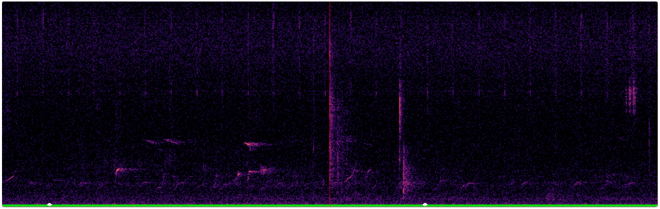
# Birdsong Detection

Spectrogram of an audio file from the forest (screenshot from the database):



# Birdsong Detection

Spectrogram of an audio file from the forest (screenshot from the database):

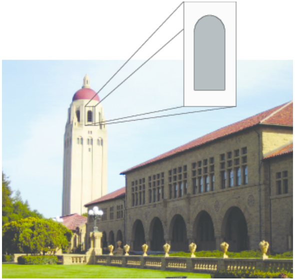


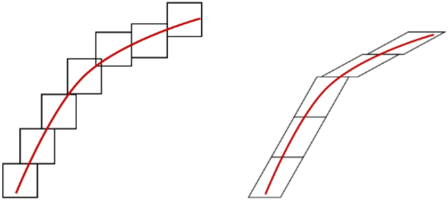
Step by step:

- Cut signal into five second chunks
- Denoise chunks (see Compression)
- Erosion with $3 \times 3$ kernel
- Are any pixels left?
- (Later) Look at classification accuracy

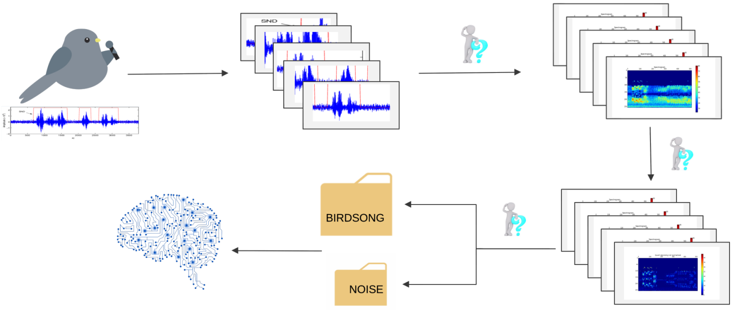
# Birdsong Detection - Outlook

Wavelet (or better: Shearlet) transform for edge detection:





# Birdsong Classification - Workflow



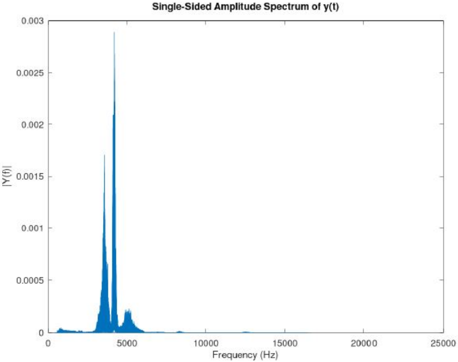
# Integration of Gabor Transform into CNN

Gabor transform (with $g^*(x) = g(-x)$):

$$V_g f(x, \omega) = \langle f, M_\omega T_x g \rangle = |(f * M_\omega g^*)(x)|$$

- Can be done in first layer of CNN
- (Later) Shearlet transform could be done in second layer
- Disadvantage: Preprocessing without spectrogram
- Advantages:
  - No need to save spectrograms
  - Classify using specific frequencies

# Classification with relevant frequencies

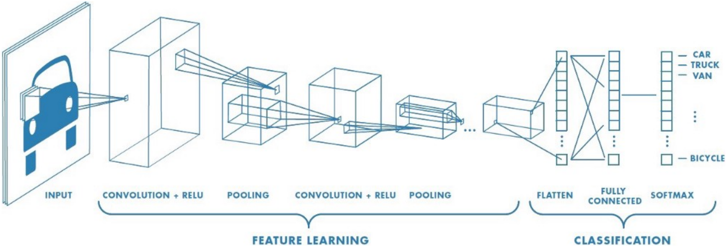



Result:

- Slightly better classification accuracy
- Faster training and classification possible

# Using an LSTM network

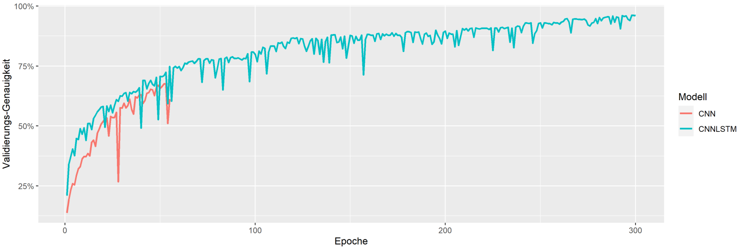
Convolutional Neural Network[3]:




- Spectrograms aren't exactly pictures
- Idea: Use time dependency by integrating LSTM layer after the feature extraction

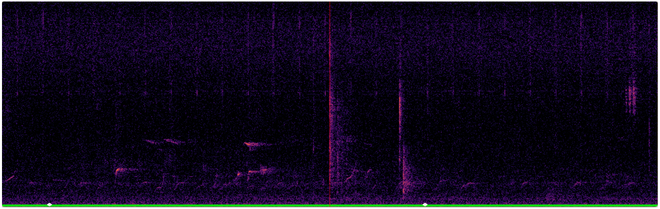
[3]https://miro.medium.com/max/1200/1*XbuW8WuRrAY5pC4t-9DZAQ.jpeg

# Classification with LSTM layer



# Outlook

Let's look at the database screenshot again:



Different signals:

- Birdsong (good time-frequency localisation)
- Crackling of twigs (more like a singularity)

Idea: Combine Gabor and Wavelet approach

# $\alpha$-Modulation Spaces

## Definition

Let $\alpha \in [0,1)$, $\varepsilon > 0$ and define

$$\begin{aligned} p_\alpha(\omega) &= \operatorname{sgn}(\omega)\left((1+(1-\alpha)\,|\omega|)^{1/(1-\alpha)} - 1\right), \\ \beta_\alpha(\omega) &= (1+|\omega|)^{-\alpha}, \\ \omega_j &= p_\alpha(\varepsilon j), \\ x_{j,k} &= \varepsilon \beta_\alpha(\omega_j) k. \end{aligned}$$

Then, we can compute the *$\alpha$-modulation coefficients*

$$c_{j,k} = \langle f, T_{x_{j,k}} M_{\omega_j} D_{\beta_\alpha(\omega_j)} g \rangle.$$

Edge cases:

- $\alpha = 0$: Gabor transform
- $\alpha \approx 1$: (Close to) Wavelet transform

# Thank you!

Extra thanks:

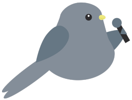
Stephan Dahlke

Pavel Tafo

Daniel Schaaf

Jacqueline Beinecke

Something for the ears:



- heuersv.de/original.wav
- heuersv.de/denoised.wav