Birdsong recognition using Gabor frames

GAMM 2021

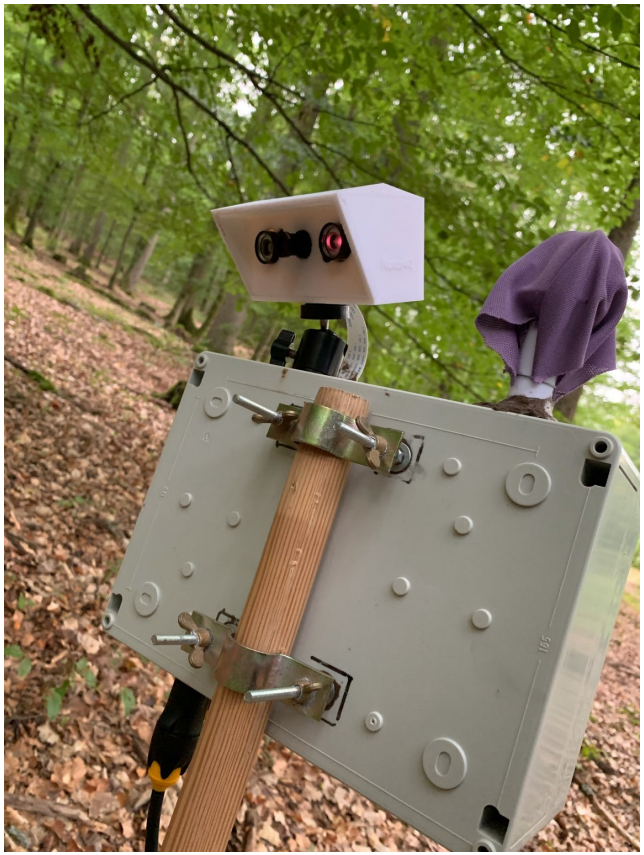
Sven Heuer

Philipps-University Marburg

Mar 18th, 2021

Motivation

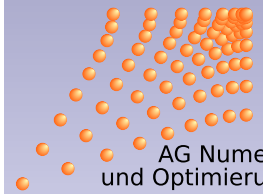
Project: Nature 4.0 | Sensing Biodiversity



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

Questions

1. How can we (denoise and) compress the raw audio files effektively?
2. How can we (detect and) classify birdsongs in the files?



Basics

Definition

Short-time Fourier transform (STFT) of f with window function g:

$$V_g f(x,\omega) = \int_{\mathbb{R}} f(t)\overline{g(t-x)}e^{-2\pi i t\omega}\mathrm{d}t, \quad x,\omega \in \mathbb{R}.$$

Basics

Definition

Short-time Fourier transform (STFT) of f with window function g:

$$V_g f(x,\omega) = \int_{\mathbb{R}} f(t)\overline{g(t-x)}e^{-2\pi i t\omega}\mathrm{d}t, \quad x,\omega \in \mathbb{R}.$$

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:

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

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

Frame reconstruction

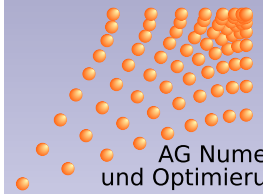
Lemma

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

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



Compression

Preconditions:

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

Theorem

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 $N = \#I_\mu$):

$$\|f - f_\mu\|_{\mathcal{L}^2}^2 \leq C_2^2 \|f\|_{\mathcal{M}^p(\mathbb{R})}^2 N^{1-2/p}.$$

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.

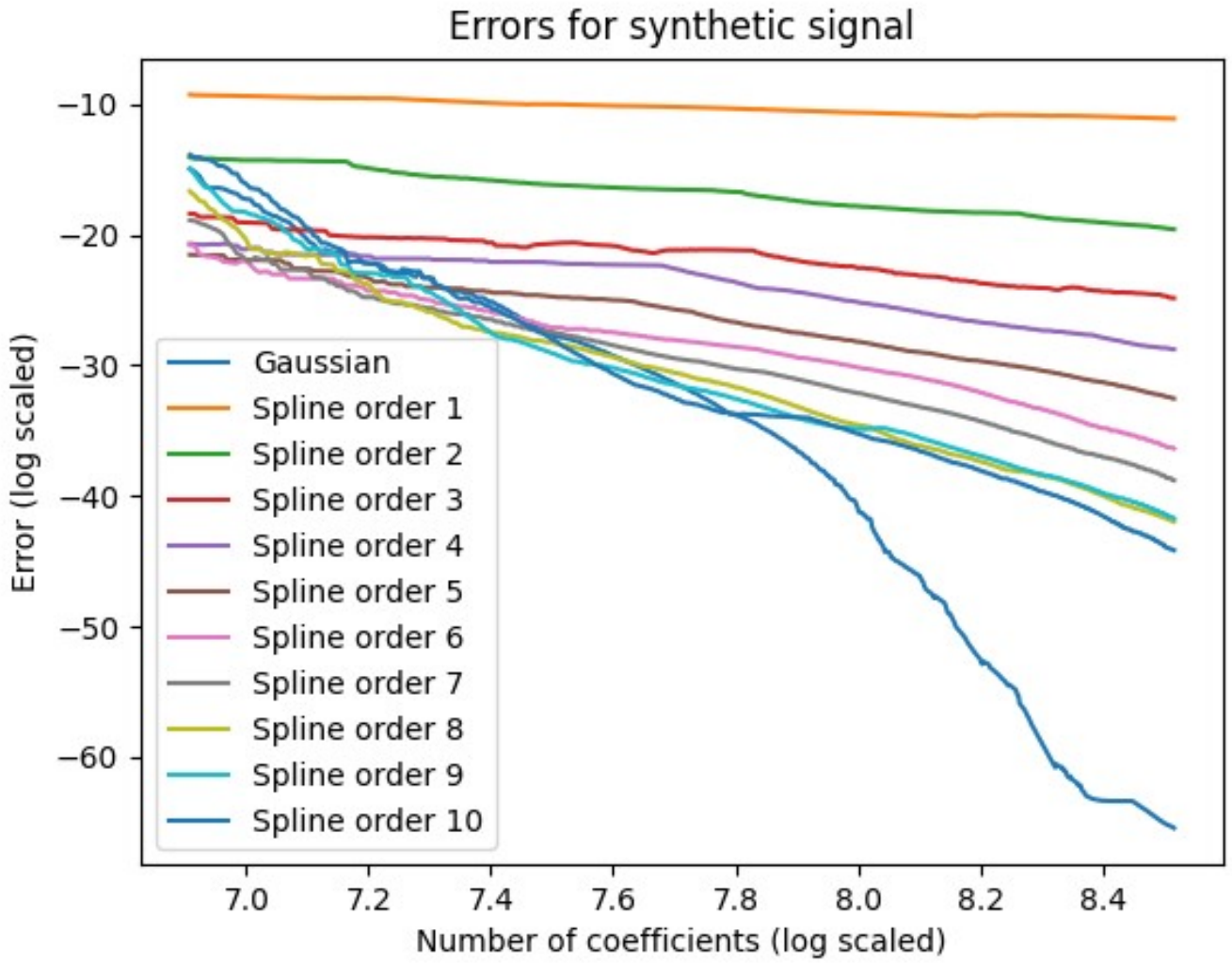
Theorem

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

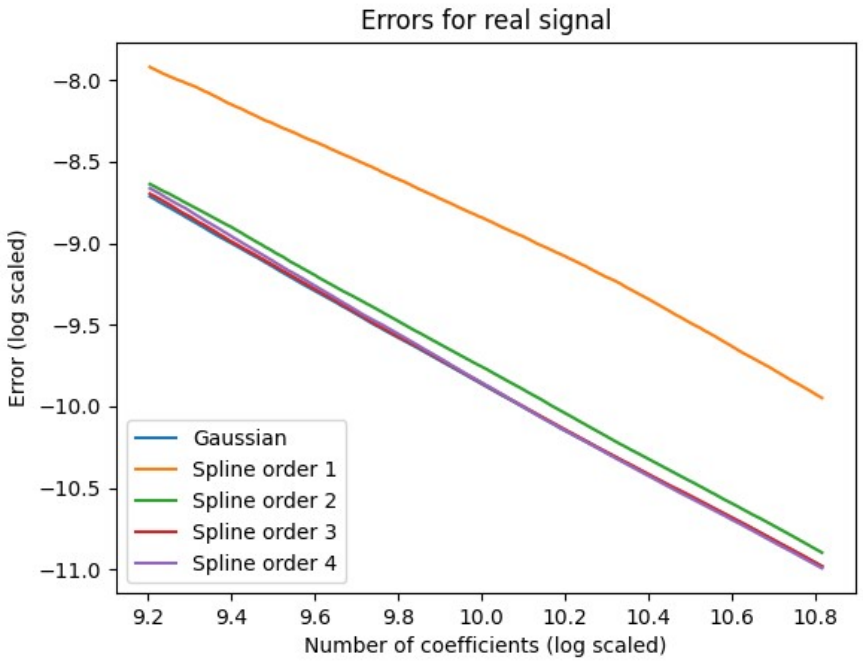
$$g \in \mathcal{M}^p(\mathbb{R})$$

for all $p > 1/k$.

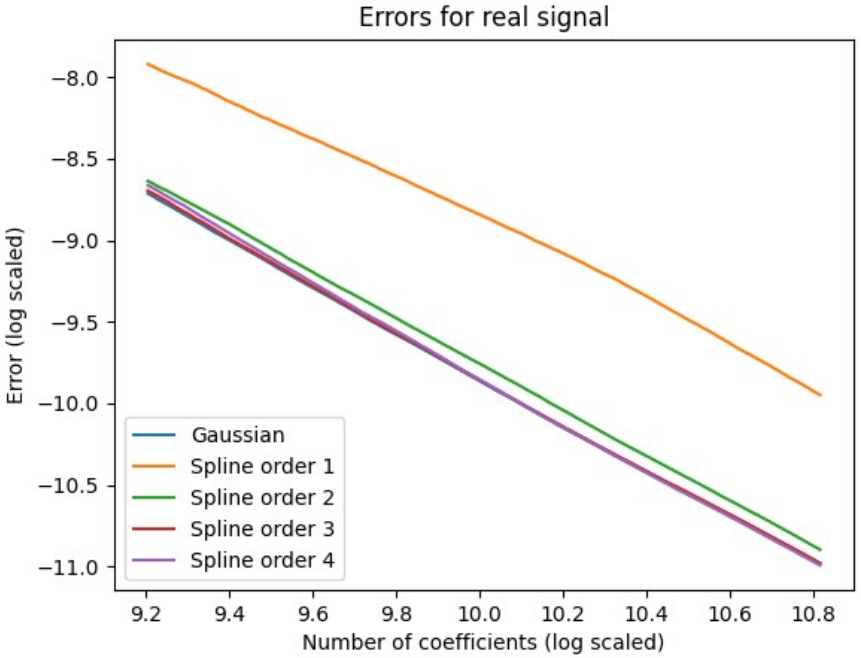
Simulations - synthetic Signal



Simulations real signal



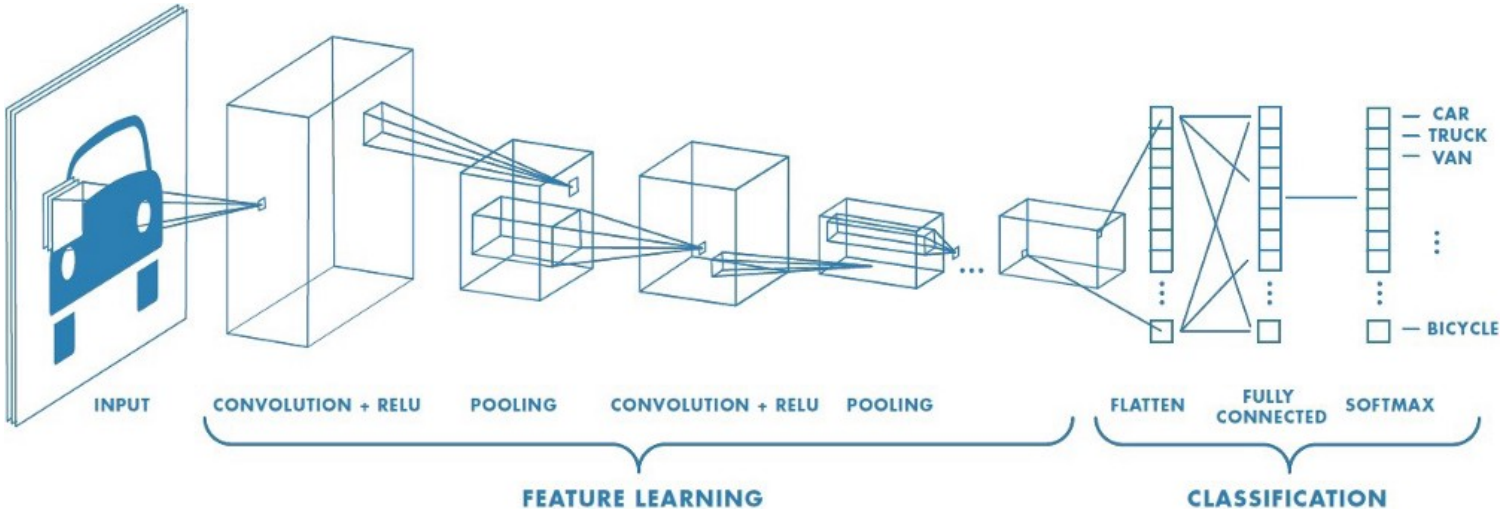
Simulations real signal



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

Classification

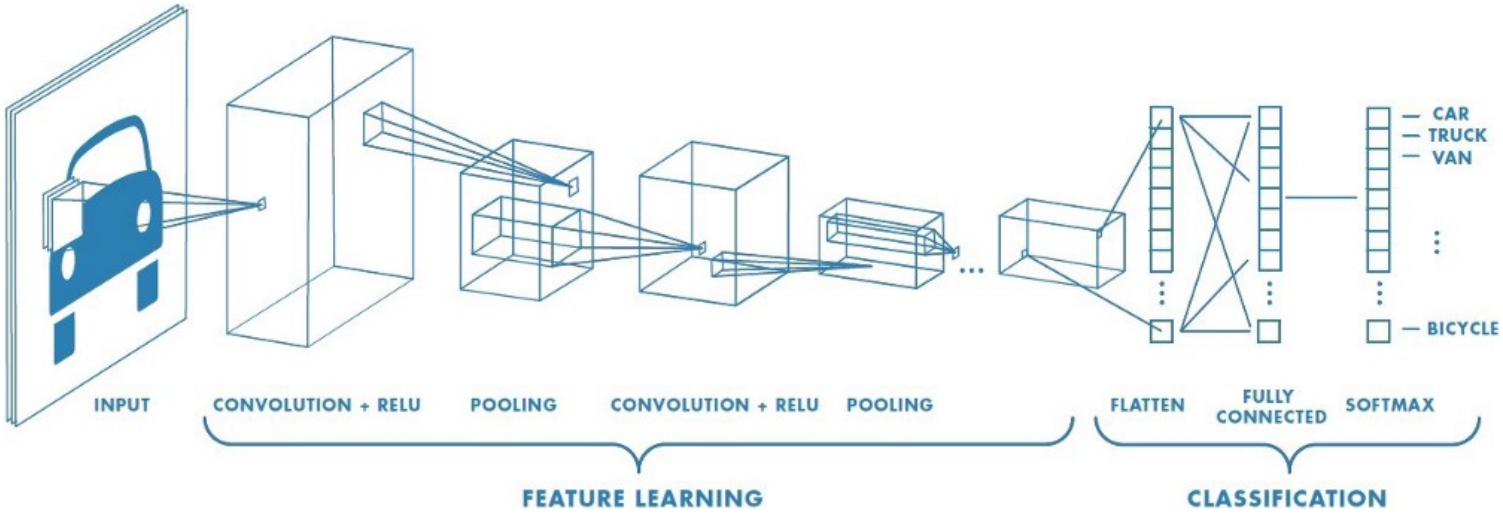
Current approach: STFT $\rightarrow$ Convolutional Neural Network[1]



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

Classification

Current approach: STFT $\rightarrow$ Convolutional Neural Network[1]

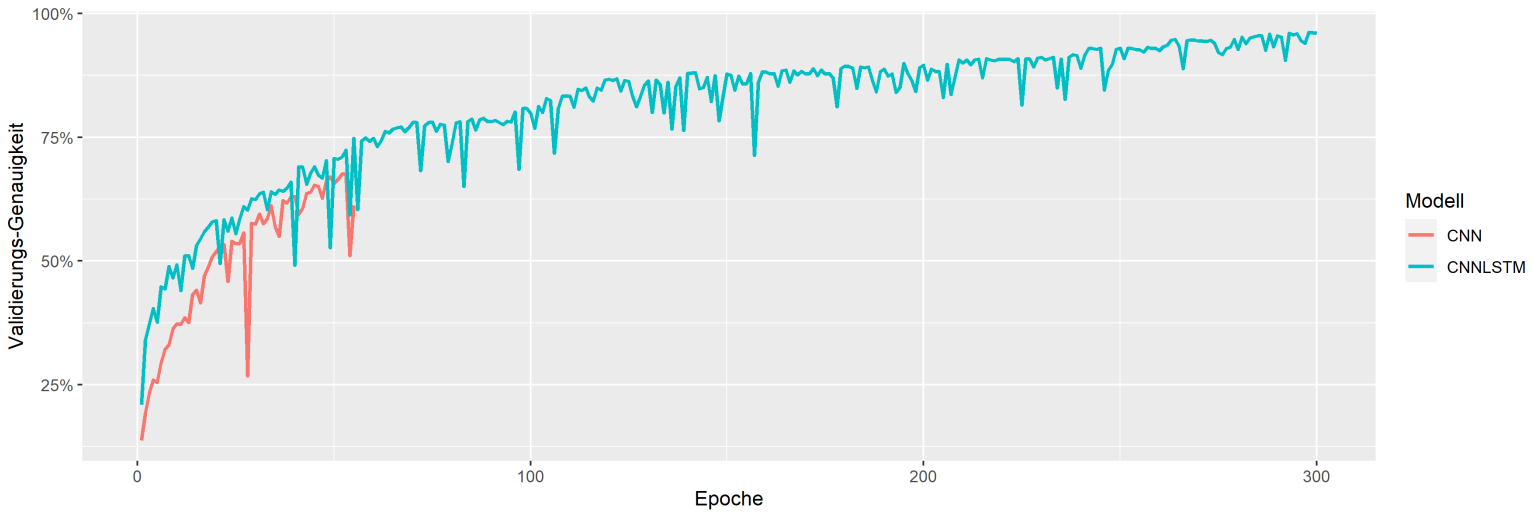


Modifications:

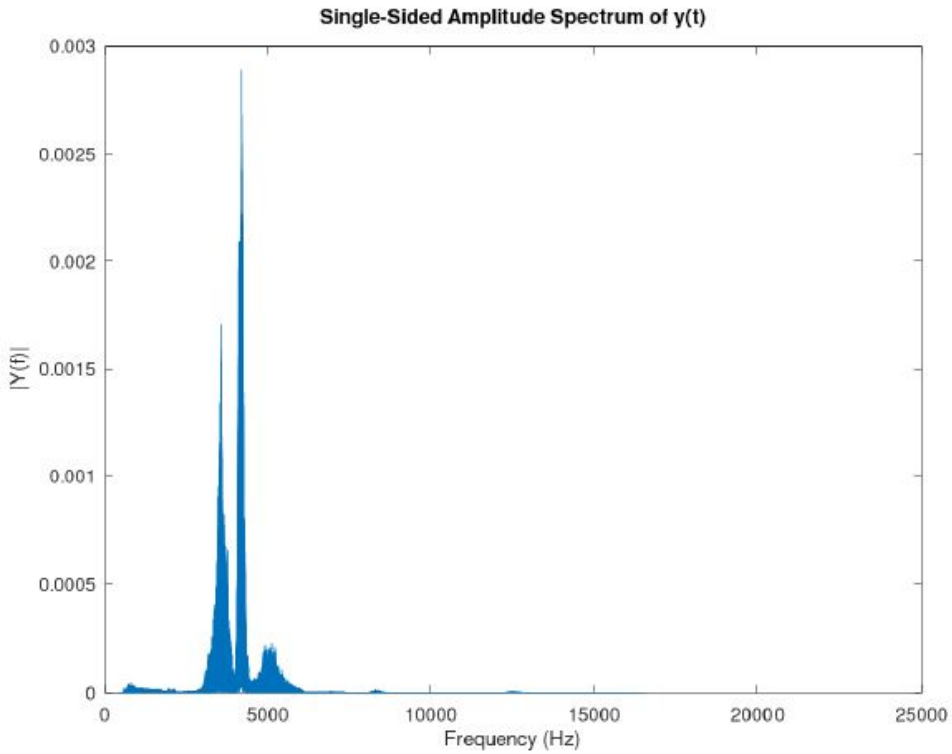
- Change the CNN structure: LSTM layer after convolutions
- Change the input: Focus on relevant frequencies

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

Classification with LSTM layer



Classification with relevant frequencies

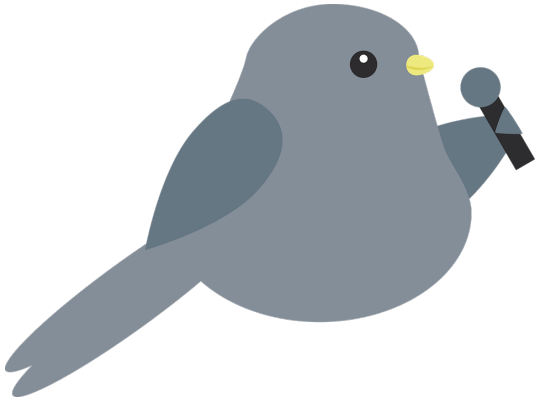


Result:

- Slightly better classification accuracy
- Faster training and classification possible

Thank you!

Something for the ears[2]:



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

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