Classification of Bioacoustic Time Series Utilizing Pulse Detection, Time and Frequency Features and Data Fusion

Christian Dietrich *, Friedhelm Schwenker *, Klaus Riede **, Günther Palm *

* Department of Neural Information Processing - University of Ulm,
D-89069 Ulm, Germany
{dietrich, schwenker, palm}@neuro.informatik.uni-ulm.de

** Zoological Institute of Scientific Research and Museum Alexander Koenig,
D-53113 Bonn, Germany
k.riede.zfmk@uni-bonn.de

ABSTRACT

Classifying the sounds of species is a fundamental challenge in the study of animal vocalizations. Most of these studies are based on manual inspection and labeling of sound spectra, which relies on agreement between human experts.

In this study recorded songs of crickets *(Grylloidea)* from Thailand and Ecuador are analysed and classified automatically. For this, the locations of pulses are determined and different features from the time and frequency domain are extracted automatically from the time series. For the categorization of the sound patterns these different features are combined through data fusion, temporal fusion and decision fusion. Local features and global features of the sound patterns are distinguished. For the classification a fuzzy-k-nearest-neighbour classifier is used. This classifier scheme exhibits a large similarity to artificial neural networks, in particular to radial basis function neural networks. We present classification results for a data set of 28 different species.

I. INTRODUCTION

Many insects produce sounds, either for defence or sexual communication. Males of crickets, katydids and grasshoppers *(Orthoptera)* "sing" by using specialised, chitinous structures on wings and legs as stridulatory apparatus [Rie98]. Crickets *(Grylloidea)* in particular are well known for their production of pure tone pulses between 2 and 11 kHz [IK98, Nis99], i.e. well within the human audible range.

The *temporal structure* is species-specific and highly organised on time scales of different orders of magnitude (see Figure 1). Besides sonograms, biologists use three features to label species-specific song parameters: frequency (2-11 kHz), pulse interval (10-150 Hz) and chirp interval (2 - 0.01 Hz, often irregular) [Ott92]. Female crickets are attracted by songs of conspecific males. This so-called *phonotaxis* has been studied extensively for several species of Gryllus spp. [HML89]. Schildberger (1994) has shown for these species the pulse repition rate is the principal feature for song recognition, and found neural correlates of female phonotactic behaviour. Grylloidea are a species-rich group comprising about 8,000 described species, and reach their highest diversity in the tropics. There are probably many more undescribed species. Recent estimates of species numbers within tropical forests exceed the number of described taxa by one order of magnitude [HKA95]. This diversity is threatened by the ongoing rapid destruction of tropical habitats and has led to the necessity for quick surveys to identify areas rich in biodiversity. Sound recordings are a valuable tool for non-invasive monitoring of singing animals, especially in tropical areas under threat. In addition, many Orthoptera are sensitive indicator species for habitat quality [Rie98].

Bioacoustic song analysis already forms an integral part within species descriptions of "new" Orthoptera [Ing97]. Cricket songs recorded from one tropical location in Ecuador could be classified qualitatetively within a parameter space of carrier frequency and pulse intervals [Rie98]. By catching voucher specimens, Nischk (1999) could show that each cluster of feature vectors is indeed produced by a distinct species. His dataset has been used for the numerical evaluation presented below (Section 6). At present, a specimen-based multimedia database is set up, bringing together data from different sound archives within the "virtual phonothek" (`http://www.dorsa.de`). Besides the mentioned features used by traditional, descriptive bioacoustics, other signal parameters can be analysed in order to classify animal sounds. Examples for such parameters are the periodogram, the energy contour and the frequency contour.

The major topic of this paper is the automated classification of crickets based on their sound patterns. The whole classification task includes *pulse detection*, *feature extraction*, *classification* and *classification fusion*. In the pulse detection previously filtered signals are used to estimate the location of the pulses on the basis of the signal's energy. These extracted pulses are used in all feature extraction steps.
In this investigation six different features are considered:

1. Frequency contour of pulses ($\boldsymbol{C}$)
2. Energy contour of pulses ($\boldsymbol{E}$)
3. Parzen density of pulse distances ($\boldsymbol{D}$)
4. Temporal structure of the pulses ($\boldsymbol{T}$)
5. Pulse length ($\boldsymbol{L}$)
6. Pulse frequency ($\boldsymbol{F}$)

We distinguish between *global features* ($\boldsymbol{C}$, $\boldsymbol{E}$, $\boldsymbol{D}$) which are defined on the whole cricket song, and *local features* ($\boldsymbol{T}$, $\boldsymbol{L}$, $\boldsymbol{F}$) which are characteristic for single pulses. Features $\boldsymbol{C}$, $\boldsymbol{E}$ and $\boldsymbol{F}$ are calculated in the *frequency domain* and the features $\boldsymbol{D}$, $\boldsymbol{T}$ and $\boldsymbol{L}$ in the *time domain*. The classification of the cricket's sounds or more precisely the classification of the feature vectors extracted from the sounds is performed through the nearest-neighbour classifier. In order to combine different features we utilize the so-called fuzzy-k-nearest-neighbour classifier.
The paper is organized as follows: In Section 2 the pulse detection algorithm is introduced; Section 3 deals with the extraction of the six different features. The k-nearest-neighbour classifiers (standard and fuzzy) are presented in Section 4. The fusion of different features and classifier decisions is shown in Section 5. In Section 6 the dataset and the numerical evaluation is presented. Finally a conclusion is given in Section 7.

II. PULSE DETECTION

The unaided human ear neither resolves the temporal structure nor the full frequency range of insect songs [Rie98]. In *automated bioacoustics* the temporal structure of the pulses is an important feature for the classification of crickets [Nis99]. Therefore the detection of pulses is an important issue for the following feature extraction and the classification task. The pulse detection algorithm performs:

1. Normalisation and filtering of the raw time signal, and
2. Signal segmentation of the filtered signal into single pulses based on the signal's energy.

The result of this pulse detection algorithm are the onsets and offsets of the pulses. These pulses are used in the further feature extraction procedures to derive the characteristic features.

A. Signal normalisation and filtering

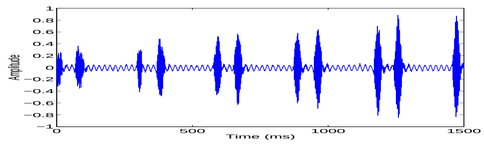
Let $s(t), t = 1, ..., T$, be the a finite signal which contains the sound pattern of a single cricket. To prevent the influence of the sound volume for the further classification the signal's amplitude is normalized through

$$\hat{s}(t) = \frac{s(t)}{||s||_\infty} \tag{1}$$

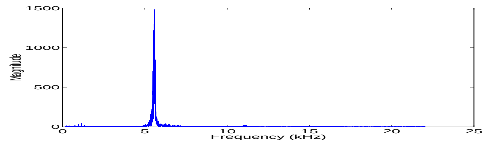
where $||s||_\infty = \max\{|s(t)| : t = 1, ..., T\}$ denotes the maximum norm.
For the signal segmentation a bandpass-filtered signal $\tilde{s}(t)$ is used to suppress the sounds of the environment. Fourier analysis is applied to calculate the filtered signal $\tilde{s}(t)$. Nowadays, in almost all software packages for signal processing an efficient **F**ast-**F**ourier-**T**ransformation (FFT) implementation is available. We use `Matlab` and a set of programs written in `C++` for the numerical experiments in this study.
Frequencies of a certain cricket are restricted to a narrow band (see Figure 1b) because the sound of single crickets is strongly resonant [CFS98]. To get rid of the sounds of the environment the signal is filtered with a highly selective bandpass-filter, the passband can be determined automatically at the frequency range of the cricket.
In the following we describe this automated filtering. First the frequency with the highest intensity is determined through

$$\omega^* = \text{argmax}_\omega |\hat{S}(\omega)W(\omega)| \tag{2}$$

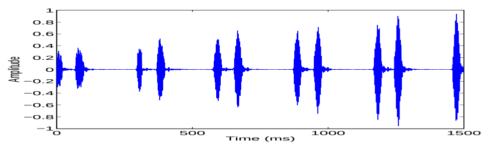
where $W(\omega)$ is a probability density function defined in the frequency range of the crickets sounds and $\hat{S}(\omega)$ is the frequency spectra of the normalized signal $\hat{s}$ (see Eq. 1).



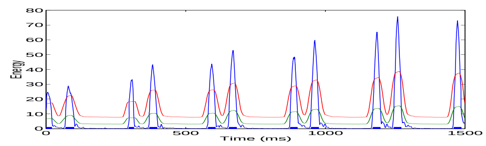
(a) The waveform $\hat{s}(t)$ after normalisation



(b) The Fourier-spectra $\hat{S}(\omega)$



(c) The filtered signal $\tilde{s}(t)$



(d) The energy functions for signal segmentation. The figure shows the thresholds $F_1(t)$, $F_\alpha(t)$ and the energy function $E_U(t)$ (see Eq. 8, 6).

Figure 1: A short "song" of a cricket from the species *Noctitrella glabra* (time window 1500 ms).

For ω^* the transfer function (or frequency response) of the bandpass-filter if defined through

$$H_\delta(\omega) = \begin{cases} 1, & |\omega| - |\omega^*| < \delta \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where 2δ is the size of the passband. This filter is used to calculate the filtered signal (see Figure 1c)

$$\tilde{s}(t) = \mathcal{F}^{-1}\{\hat{S}(w)H_\delta(\omega)\} \tag{4}$$

where $\mathcal{F}^{-1}\{\cdot\}$ denotes the inverse Fourier transform. The filtered signal $\tilde{s}(t)$ is limited to the frequency range $[\omega^* - \delta, \omega^* + \delta]$ which is assumed to contain the sounds of a cricket. After normalisation and filtering the signal $s(t)$ with $H_\delta(\omega)$ the filtered signal $\tilde{s}(t)$ is free from background sounds, such as the voices of birds.

B. Signal segmentation through the signal's energy

An important problem in signal processing is the detection of the presence of a signal within a background of noise. This problem is often referred to as the *end point location problem*

[RS97]. In speech recognition systems end point detection is used to get the relevant parts of the speech signal where the speaker is active. We use end point detection to locate the cricket's pulses.

Rabiner and Sambur 75 published an end point algoritm for speech segmentation using the signal-energy and the zero-crossing-rate of the signal to localize the onset and offset of an utterance. This algorithm can be used in almost any background environment with a signal-to-noise ratio of at least 30 dB. The differences between human speech and animal vocalizations, and the different conditions under which they are recorded, are significant and have to be taken into account [KM98]. For pulse segmentation we use an algorithm which determines the onset and offset of the pulses of cricket songs. The algorithm is similar to that of Rabiner and Sambur, but with significant modifications, particularly for the threshold function.

For the pulse detection algorithm the total energy E of the previously filtered signal $\tilde{s}$ is defined by

$$E = \sum_{\tau=-\infty}^{+\infty} |\tilde{s}(\tau)| \tag{5}$$

and the *short time energy* of the signal $\tilde{s}$ inside a window $\mathcal{W}_t^U$ is given through

$$E_U(t) = \sum_{\tau=-\infty}^{+\infty} |\tilde{s}(\tau)| \mathcal{W}_t^U(\tau) \tag{6}$$

where

$$\mathcal{W}_t^U(\tau) = \begin{cases} 1, & \tau \in [t - \frac{U}{2}, t + \frac{U}{2}] \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

is a centered rectangular window of size U at time t. A threshold-function $F_a(t)$ is used to determine the pulses of the cricket's sound. In speech recognition this threshold $F_a(t)$ is typically a constant function [RS97].

It can be observed that for some species the short time energy function $E_U(t)$ increases inside chirps. In Figure 2 an energy function $E_U(t)$ is shown where the minimal energy between two pulses gets higher during a chirp. Therefore, a dynamic threshold-function is applied:

$$F_a(t) = \theta + \frac{a}{\xi} \sum_{\tau=-\infty}^{+\infty} |\tilde{s}(\tau)| \mathcal{W}_t^V(\tau) = \theta + \frac{a}{\xi} E_V(t) \tag{8}$$

with parameter $a > 0$, rectangle window size $V > U$ (see Eq. 7) and scaling parameter $\xi > \frac{V}{U}$. Such a threshold-function is depending on the short time energy of the filtered time signal $\tilde{s}$ in a finite time window $\mathcal{W}_t^V$.
The silence energy E_{min} and the peak energy E_{max} given through

$$E_{max} = max_t E_U(t)$$
$$E_{min} = min_t E_U(t)$$

are used to define the minimal threshold θ (see Eq. 8) which is set to

$$\theta = E_{min} + \gamma(E_{max} - E_{min}) \tag{9}$$

with $0 < \gamma < 1$.

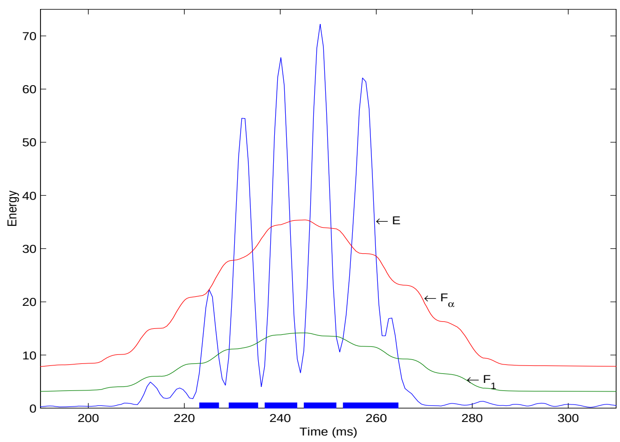


Figure 2: Detection of 5 pulses inside a single chirp of the species *Zvenella transversa*. The Figure shows the thresholds $F_1(t)$, $F_\alpha(t)$ and the energy function $E_U(t)$

Our algorithm uses two threshold-functions: the *lower energy threshold* $F_1(t)$ and the *upper energy threshold* $F_\alpha(t)$ with $\alpha > 1$, obviously $F_1(t) < F_\alpha(t)$. The estimation of the onsets through the energy function $E_U(t)$ (see Eq. 6) and the thresholds $F_1(t)$ and $F_\alpha(t)$ (see Eq. 8) is accomplished by Algorithm 1.

```
τ = 0; p = 0 A = 1
for t = V/2, ..., T − V/2 step Δt
    if (E_U(t) ≥ F_1(t))
        if (τ > 0)
            if (A = 1)
                if (E_U(t) ≥ F_α(t))
                    p = p + 1; λ_p = OS; A = 0
                else
                    if (τ > ψ) τ = 0
                    else τ = τ + Δt end
                end
            end
        else
            τ = τ + Δt; OS = t
        end
    else
        τ = 0; A = 1;
    end
end
```

Algorithm 1: Searching the onsets λ_p of the energy function $E_U(t)$.

To detect an onset of a pulse the short time energy $E_U(t)$ has to exceed both threshold functions $F_1(t)$ and $F_\alpha(t)$ within a specified time interval ψ. A similar algorithm is used to estimate the offset locations of the pulses.
In the numerical experiments the pulse segmentation algorithm is used with the parameters listed in Table I. It should be noticed that these parameters are completely different to the parameters in speech recognition systems. For example, the window size $U = 2.7$ ms is much smaller than the windows in speech recognition systems (10 ms) [RJ93].
The results of the signal segmentation algorithm is n the number of pulses in the signal and the sequences of onsets $\lambda = (\lambda_1, ..., \lambda_n)$ and offsets $\mu = (\mu_1, ..., \mu_n)$ of the detected pulses.

III. FEATURE EXTRACTION

For the classification of the cricket songs the sounds must be transformed into feature vectors suited for the pattern recognition task. A pattern recognition system can be considered as a two stage system: the extraction of features from the time signal and the classification of the feature vectors [BJ92]. In the feature extraction different *feature detectors* are defined in order to extract a set of characteristic signal properties which can be used as input features for the automatic classifier system.
In principle there are two approaches of feature extraction for time series:

1. **Global features**. These features based on global characteristics or information of the whole time series (see Section B., C. and D.).

2. **Local features**. These local features are derived from subsets of the whole time series, which are usually determined by local time windows. In the context of this paper the time windows are located at the detected pulses. Moving the window over the whole time series leads to a sequence of feature vectors (see Section E., F. and G.).

A. Sonograms of individual pulses

Sonograms may not be the best input features for classifying with artificial neural networks or other statistical classifier schemes [MMR98], because the number of features is too high. But they allow accurate measurement of the durations and frequencies of notes and illustrate the way in which the frequency and amplitude of a sound change in time [EE94]. Therefore, sonograms may be a good basis to extract features like the frequency and the energy contour of a sound pattern, e.g. a chirp or a pulse.
We distinguish two types of sonograms: sonograms with a *fixed time-resolution* and sonograms with a *fixed number of sampling points.*
Let $s(t)$ be the signal which contains the sound of a cricket. Then the amplitude signal of the k-th pulse is given by

$$v_k(t) = \begin{cases} s(t), & t \in [\lambda_k, \mu_k] \\ 0, & \text{otherwise} \end{cases}. \tag{10}$$

For simplicity we set $v := v_k$, $\mu := \mu_k$ and $\lambda := \lambda_k$. Then the sonogram of a single pulse v

is calculated by m short time Fourier transformations defined as

$$E_{lj} = \sum_{\tau=-\infty}^{+\infty} v(\tau)\mathcal{W}_{t_j}^{V}(\tau)e^{-i\frac{2\pi}{V}\tau l}, \quad j = 1, ..., m \tag{11}$$

where $l = 0, ..., \frac{V}{2}$ is the frequency band [RJ93]. Here $\mathcal{W}_{t_j}^{V}$ is a rectangular window of size V (see Eq. 7) and the sampling vector $T = (t_1, ..., t_m)$ is given by

$$\begin{aligned} t_1 &= \lambda - (\frac{1}{2} - \alpha)V \\ t_j &= t_1 + (j-1)\beta V \end{aligned} \tag{12}$$

$j = 2, ..., m$. The overlap between two consecutive sampling windows is $(1-\beta)V$ and αV is the overlap between the first window $\mathcal{W}_{t_1}^{V}$ and the time signal v (see Figure 3).

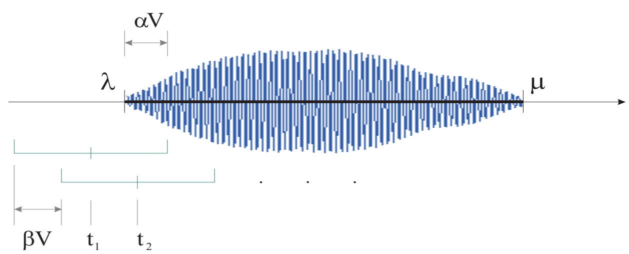


Figure 3: Sampling Windows to extract frequency spectra from a single pulse

The length of the sampling vector T is then given by

$$m = \left\lceil \frac{\mu - \lambda + (1-\alpha)V - \alpha V}{\beta V} \right\rceil . \tag{13}$$

This is the common type of sonogram where the number of spectra depends on the length of the analyzed signal part.

Sonograms with a constant number of spectra are of interest in order to extract feature vectors of constant length m. These sonograms can be calculated by modifying the sampling vector T. For window size V and signal length $l = \mu - \lambda$ two cases have to be distinguished (see Eq. 12 and Figure 3):

1. $mV \leq l$
   Here α is set equal to 1 and $\beta = 1 + \frac{l-mV}{(m-1)V}$.

2. $mV > l$
   Here the parameters α and β are set to $\alpha = \beta = 1 - \frac{mV-l}{(m+1)V}$.

B. Frequency contour of pulses

The frequency contour of vocalisations is a signal parameter which is often used to classify sounds of animals. Murray et al. 98 extracted the frequency contour by *fundamental frequency analysis* in order to classify whistles of killer whales. One method to extract the frequency contour f from a discrete spectra is to determine the band of the frequency with maximal energy [VHH98].

For a sonogram with fixed length m of a single pulse, let ω_l be the mean frequency of the l-th frequency band and E_{lj} be the energy of the l-th frequency band within the j-th time window (see Eq. 11). Then

$$l_j^* = \mathrm{argmax}_l E_{lj} \tag{14}$$

is the frequency band with the maximal energy. Including adjacent frequency bands and time windows is sharpening the result. Therefore we do not use $\omega_{l_j^*}$ the frequency band with the maximal energy but the weighted average

$$f_j := \sum_{l=l_j^*-\rho}^{l_j^*+\rho} \sum_{k=j-\sigma}^{j+\sigma} \frac{\omega_l E_{lk}}{\sum_{l=l_j^*-\rho}^{l_j^*+\rho} \sum_{k=j-\sigma}^{j+\sigma} E_{lk}}. \tag{15}$$

Here the parameters ρ and σ determine the size of the averaging window in the frequency and time domain.

The frequency contour of a single pulse is then given by $C = (f_1, ..., f_m) \in \mathbb{R}^m$. Now we assume a sequence of n pulses, each represented through a sonogram of fixed length m. Then we denote with $C^k \in \mathbb{R}^m$ the frequency contour of the k-th pulse. The characteristic contour C^{i^*} is defined as the frequency contour with the minimal Euclidean distance to all other contours C^h.

$$i^* = \mathrm{argmin}_i \{\sum_{h=1}^{n} ||C^h - C^i|| : i = 1, ..., n\} \tag{16}$$

The vector $C^{i^*} \in \mathbb{R}^m$ is used as feature for the classification. For the parameters used in the numerical evaluation see Table I.

C. Energy contour of pulses

For the extraction of the energy contour the spectral energys E_{lj} can be used. In difference to the frequency contour $C \in \mathbb{R}^m$ the energy contour is extracted from sonograms with a fixed time resolution. Using the parameters α and β (see Eq. 11 and 12) as defined in

Table I leads to a time resolution of 1.3 ms (60 samples). Then the total energy of the j-th spectrum E_j (see Eq. 11) is given by

$$E_j = \sum_{l=1}^{m} E_{lj}. \tag{17}$$

Again we assume a sequence of n pulses, each represented through a sonogram where the number of spectra depends on the pulse length. Then we denote $E^k = (E_1^k, ..., E_{m_k}^k)$ as the energy course of the k-th pulse where m_k defines the number of spectra of this pulse. To extract feature vectors of length d we set

$$E^k = \begin{cases} (E_1^k, ..., E_d^k) \in \mathbb{R}^d, & d \leq m_i \\ (E_1^k, ..., E_{m_i}^k, 0, ..., 0) \in \mathbb{R}^d, & d > m_i. \end{cases} \tag{18}$$

The discrete feature vector E for the classification is the mean of the energy courses $E^1, ..., E^n$ of the single pulses (see Figure 4d).

D. Parzen density of pulse distances

To determine a probability density function over the distance between two consecutive pulses a histogram procedure is used [Par62]. Let $\lambda = (\lambda_1, ..., \lambda_n)$ be a sequence containing the onsets of the pulses. Then

$$\Delta_j = \lambda_{j+1} - \lambda_j, \quad j \in \{1, ..., n-1\} \tag{19}$$

is the distance between the j-th and the $(j+1)$-th pulse. These distances are used to estimate a one dimensional density function D_Δ based on the Gaussian density as kernel function [KB98]

$$f_{\Delta_j}(t) = e^{-\frac{(t-\Delta_j)^2}{2\sigma^2}}, \tag{20}$$

with variance $\sigma^2 > 0$. The Gaussian may be used under the assumption that the estimated density function D_Δ is continuous [Rus88]. Then the density function D_Δ is given by

$$D_\Delta(t) = \frac{1}{n-1} \sum_{j=1}^{n-1} e^{-\frac{(t-\Delta_j)^2}{2\sigma^2}}. \tag{21}$$

To approximate a discrete feature vector the function D_Δ is sampled with linear increasing time steps $t = (t_1, ..., t_m)$:

$$t_i = \gamma t_{i-1} + \delta, \quad i = 1, ..., m \tag{22}$$

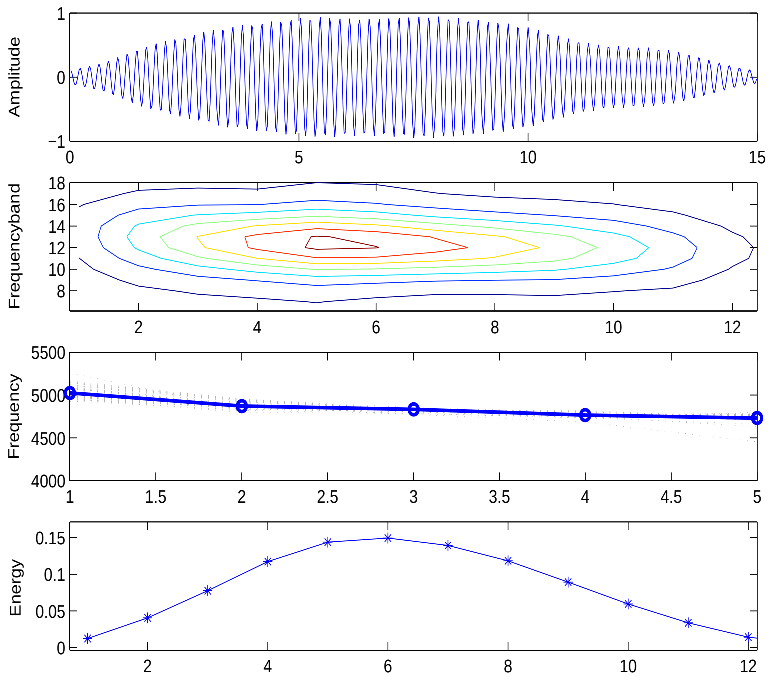


Figure 4: Some features extracted from a single pulse of the species *Noctitrella plurilingua*: a) time signal, b) sonogram as contour plot, c) individual pulse frequencies (small dots) and their characteristic frequency (bold line), d) energy course

from $t_0 = 0$ to t_m. Here m is determined in such a way that t_m is approximately 500 ms depending on the parameters of γ and δ (see Table I and Figure 6).

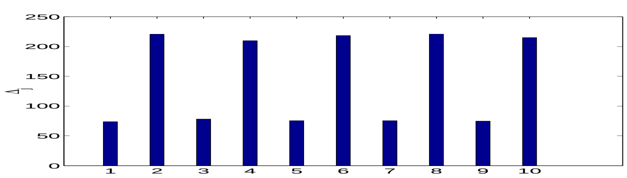


Figure 5: Distances between pulses of the *Noctitrella glabra* (signal length 1.5 sec). Each bar shows the distance between two adjacent pulses Δ_j.

E. Temporal structure of pulses

The basis to extract these classification vectors are the distances between pulses $\Delta = (\Delta_1, ..., \Delta_{n-1})$ (see Eq. 19). These features are extracted using a $d-$tuple encoding scheme producing $n - d$ feature vectors $T_i \in \mathbb{R}^d$

$$T_i := (\Delta_i, \Delta_{i+1}, ..., \Delta_{i+d-1}) \in \mathbb{R}^d, \quad i = 1, ..., n - d. \tag{23}$$

which are used for the automated classification.

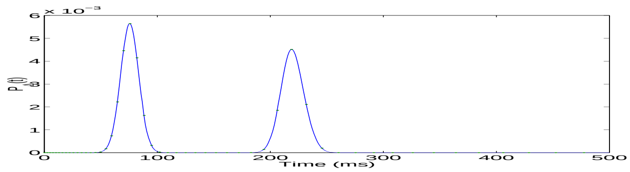


Figure 6: The probability density function of the pulse distances of the species *Noctitrella glabra* shows clusters of pulse distances with centers at 74 ms and 220 ms (see also Figure 5).

F. Pulse length

The mean pulse length L_i for a set of d adjacent pulses is determined from the onsets λ and offsets μ of single pulses

$$L_i = \frac{1}{d} \sum_{j=i}^{i+d-1} (\mu_j - \lambda_j), \quad i = 1, ..., n - d. \tag{24}$$

G. Pulse frequency

Let $f^k \in \mathbb{R}^m$, $k = 1, ..., n$ be the frequency contour of the k-th pulse and $\bar{f}^k = \frac{1}{m} \sum_{j=1}^{m} f_j^k$ be the average frequency of this pulse. Then the average frequency over d pulses is

$$F_i = \frac{1}{d} \sum_{j=i}^{i+d-1} \bar{f}^k, \quad i = 1, ..., n - d \tag{25}$$

which is used as input for the classifier.

IV. CLASSIFICATION WITH THE FUZZY-K-NEAREST-NEIGHBOUR RULE

Nischk 99 indicated that cricket songs can be categorized. He showed that clusters can be detected in the parameter space made up by the average length of the interval between two pulses and the carrier frequency of the sound.

One of the most elegant and simplest classification techniques is the *k-nearest-neighbour rule* [Fuk90] which is used to classify feature vectors $x \in \mathbb{R}^d$. The classifier searches at first for the k nearest neighbours among a set of m prototypes for an *input vector* x. The k nearest neighbours are calculated utilizing the L_p-norm ($p \in [1, \infty)$, often the Manhattan distance ($p = 1$) or the Euclidean distance ($p = 2$) is used)

$$d_j^p(x, x^j) = ||x - x^j||_p = \Big(\sum_{i=1}^{d} |x_i - x_i^j|^p\Big)^{\frac{1}{p}}. \tag{26}$$

For the distances d_j^p there is a sequence $(\tau_i)_{i=1}^m \subset \{1, ..., m\}$ with the property $d_{\tau_1}^p \leq ... \leq d_{\tau_m}^p$. The k nearest neighbours of x are defined by

$$\mathcal{N}_k(x) := \{x^{\tau_1}, ..., x^{\tau_k}\} \tag{27}$$

with $k \leq m$. Let $c_i = c(x_{\tau_i})$, $i = 1, ..., k$, be the classes of the k nearest neighbours and

$$\mathcal{N}_k^j(x) = \{y \in \mathcal{N}_k(x) | c(y) = j\} \tag{28}$$

be the subset of nearest neighbours of class j. Then the classification for input x is defined through the majority:

$$j^* = \mathrm{argmax}_{j=1,\ldots,l} |\mathcal{N}_k^j(x)| \tag{29}$$

where j^* is the class which occurs most often among the k nearest neighbours. Here l is the number of classes.

To determine the class membership of an input vector x to all l classes, a fuzzy-k-nearest-neighbour classifier is used [Sin98]. Such a fuzzy classifier Δ is a mapping $\Delta : \mathbb{R}^d \to [0, 1]^l$, i.e., the output $\Delta(x) = (\Delta_1(x), ..., \Delta_l(x))$ contains the membership of x to each class. Let

$$\delta_j(x) = \frac{1}{\sum_{x^i \in \mathcal{N}_k^j(x)} ||x - x^i||_p + \alpha} \tag{30}$$

be the support for the hypothesis that j is the true label of x. We set $\delta_j(x) = 0$ if $\mathcal{N}_k^j(x) = \emptyset$. The parameter $\alpha > 0$ is used to grade low values of $||x - x_i||_p$.

After normalisation by

$$\Delta_j(x) := \frac{\delta_j(x)}{\sum_{i=1}^{l} \delta_i(x)} \tag{31}$$

we have $\Delta_j(x) \in [0, 1]$ and $\sum_{j=1}^{l} \Delta_j(x) = 1$ and call the classifier outputs *soft labels* [Kun00]. The parameter k determines the fuzzyness of the classification result. The fuzzy-k-nearest-neighbour classifier shows a similarity to artificial neural networks, in particular to competitive neural networks and radial basis function networks [PG90, SKP01].

V. FUSION

The individual features alone are not sufficent for a proper classification of the cricket song (see Table III). Furthermore the number of local feature vectors T_i, L_i and F_i, $i = 1, ..., n-d$ (see Eq. 23, 24 and 25) derived from the individual pulses depends on n of the number of pulses. So fusion can be used to combine these features. In the following three different fusion methods are distinguished

1. **Data fusion** is the combination of a set of different feature vectors by concatenation these feature vectors into one feature vector.

2. **Decision fusion** is the combination of a set of decisions based on different feature detectors into one decision through a fusion function.

3. **Temporal fusion** is the combination of a sequence of decisions calculated in different parts of the time series into a decision. The same fusion functions as for decision fusion may be used.

The decisions of a set of classifiers are used to build the *decision profile* which is a matrix of soft labels

$$DP(x) = \begin{bmatrix} \Delta_1^1 & \dots & \Delta_j^1 & \dots & \Delta_l^1 \\ \dots & \dots & \dots & \dots & \dots \\ \Delta_1^i & \dots & \Delta_j^i & \dots & \Delta_l^i \\ \dots & \dots & \dots & \dots & \dots \\ \Delta_1^o & \dots & \Delta_j^o & \dots & \Delta_l^o \end{bmatrix} \tag{32}$$

where o determines the number of classifiers and Δ^i is the output of the i-th classifier. The classification fusion may be applied through several fusion mappings e.g. *symmetrical probabilistic fusion* or *average fusion* [KB98]. Symmetrical probabilistic fusion is a straightforward approach to combine the classification results applying the Bayes' rule [Pea88] under the assumption that the classification results are independent. The symmetrical probabilistic fusion [TB96] is defined by

$$\tilde{\Delta}_j(x) = 1 - \big(1 + \alpha \left(\frac{1-P_j}{P_j}\right)^{o-1} \prod_{i=1}^{o} \frac{\Delta_i^j}{1-\Delta_i^j}\big)^{-1}. \tag{33}$$

Here P_j denotes the prior probability for class j and α is a normalizing constant. The degree of fuzzyness and the certainty of the single classifiers is important utilizing this function, because the product is zero if just a single decision Δ_i^j is zero.

The average of the classification results is given by

$$\tilde{\Delta}_j(x) = \frac{1}{o}\sum_{i=1}^{o} \Delta_j^i. \tag{34}$$

VI. NUMERICAL EVALUATION

A. Dataset

In this section we present results achieved by testing the algorithms on a dataset containing sound patterns from 28 different species from Thailand and Ecuador.The dataset consists of recordings from 108 different animals with 3 or 4 recordings per species. The cricket songs from Thailand were recorded by Ingrisch [IK98], the recordings from Ecuador were used in the doctoral thesis of Nischk 99. The sound patterns are stored in the standard WAV-format (44.100 Hz sampling frequency, 16 Bit sampling accuracy) which is also used for audio CDs.

B. Results

Because of the limited data we apply the k-fold cross-validation method [Bis95] to evaluate the proposed classification algorithms. In the k-fold cross-validation testing procedure the dataset is divided into k disjoint subsets. Then the classifier is trained k times, each time using a version of the dataset omitting exactly one of the k subsets. The omitted subset is then used to test the trained classifier. Finally the achieved classification results are averaged over all k classifier tests. In the numerical evaluation presented in this paper the cross-validation procedure has been performed for $k = 4$ using exactly one record per species for the classification test and the rest for the classifier training. In Table II, III and IV and Figure 8, 10 and 12 the calculated means of the k cross-validation testings are given.

The parameters used for the pulse segmentation and feature extraction algorithms in the classifier evaluation of this study are given in Table I.

Table I: The parameters for the pulse segmentation (upper part) and the feature extraction (lower part).

Algorithm	**Parameters**
Pulse segmentation	$\alpha = 1.4$, $\gamma = 0.05$
	$U = 2.7$ ms (120 samples)
	$V = 10.8$ ms (480 samples)
	$\beta = \frac{U}{4}$, $\xi = 32$
Sonograms	$V = 5.4$ ms (240 samples)
Frequency contour	$m = 5$, $\rho = 5$, $\sigma = 0$
Energy contour	$d = 25$, $\alpha = \beta = \frac{1}{4}$
Parzen density	$\gamma = 1.05$, $\delta = 0.2268$ ms
Temporal structure	$d = 5$
Pulse length	$d = 5$
Pulse frequency	$d = 5$

The classification based on global and local features and their combination has been investigated. First results of the extracted global features are presented. In Table II the classification error rates of the cross-validation procedure are shown for the Parzen density of pulse distances ($\boldsymbol{D}$), the frequency contour of pulses ($\boldsymbol{C}$), and the energy contour of pulses ($\boldsymbol{E}$). In order to illustrate the influence of k (the number of the nearest neighbours of the fuzzy-k-nearest-neighbour classifier) which is taken into account for the classifier decision, classification results for k between $k = 1$ and $k = 50$ are given. The parameter k determines the fuzzyness of the classifier output (see Eq. 27).

Table II: Classification results for the individual global features (upper part) and the global features combined through data fusion (lower part).

Global features	Classification error in %									
parameter k	1	3	5	7	10	15	20	30	40	50
Parzen Density ($\boldsymbol{D}$)	45.4	52.8	55.6	54.6	55.6	54.6	54.6	53.7	53.7	53.7
Frequency contour ($\boldsymbol{C}$)	60.2	62.0	59.3	65.7	63.0	63.9	63.0	63.9	63.9	63.9
Energy ($\boldsymbol{E}$)	57.4	51.9	58.3	61.1	61.1	62.0	63.0	60.2	60.2	60.2
$\boldsymbol{DC}$	36.1	48.1	48.1	44.4	52.8	50.0	50.0	49.1	49.1	49.1
$\boldsymbol{DE}$	34.3	41.7	45.4	49.1	46.3	47.2	46.3	43.5	43.5	42.6
$\boldsymbol{CE}$	37.0	40.7	38.9	41.7	43.5	40.7	40.7	40.7	40.7	40.7
$\boldsymbol{DCE}$	35.1	34.3	43.5	43.5	41.7	45.4	44.4	41.7	40.7	40.7

For each of the three features ($\boldsymbol{D}$, $\boldsymbol{C}$ and $\boldsymbol{E}$) the observed error rates are rather high (in the range of $45-65\%$ depending on k and the feature type). The classifier performance of each global feature ($\boldsymbol{D}$, $\boldsymbol{C}$, $\boldsymbol{E}$) is improved through a combination of these global features through concatenation of the single feature vectors (data fusion). But still the classification error rates for the feature pairs ($\boldsymbol{DC}$, $\boldsymbol{DE}$, $\boldsymbol{CE}$) and the feature triplet ($\boldsymbol{DCE}$) are rather high. This phenomenon can be observed in more detail in the confusion matrices of the classifiers.
Figure 8:

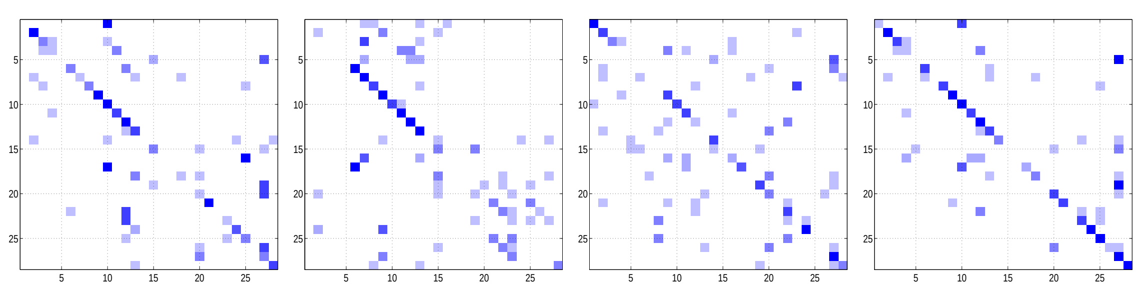


Figure 8: Mean of the confusion matrices of the cross validation runs $\boldsymbol{D}$) Parzen density function, $\boldsymbol{C}$) characteristic frequency contour, $\boldsymbol{E}$) energy contour within pulses, global features combined through data fusion.

For example in Figure 8 the confusion matrices of the fuzzy-k-nearest neighbour classifiers (with $k = 10$) are given for the three single global features (column 1: Parzen density of pulse distances ($\boldsymbol{D}$), column 2: frequency contour of pulses ($\boldsymbol{C}$) and column 3: energy contour of pulses ($\boldsymbol{E}$)) and for the combined feature triplet (column 4: $\boldsymbol{DCE}$). Here, positive confusion frequencies between individuals of different species are represented as pixels with large grey values outside the diagonal of the confusion matrix. All four confusion matrices show lots of positive confusion frequencies between different species.

The overall classification based on local features is calculated through temporal fusion by averaging all local soft classifier decisions over the whole sound pattern, see Eq. 34. The combination of local features is again calculated through data fusion, e.g. the concatenation of the single feature vectors. On the combined features the local decision is calculated through the fuzzy-k-nearest-neighbour-classifier, and then these soft decisions are averaged over the whole sound pattern (temporal fusion), see Eq. 34. The classification results for the single local features pulse length ($\boldsymbol{L}$), pulse frequency ($\boldsymbol{F}$) and temporal structure of pulses ($\boldsymbol{T}$) and for the combined features ($\boldsymbol{LF}$, $\boldsymbol{LT}$, $\boldsymbol{FT}$, $\boldsymbol{LFT}$) are given in Table III, again for $k = 1$ to $k = 50$.

Table III: Classification results of the local features after temporal fusion (upper part) and the local features combined through data fusion (lower part).

Local features	Classification error in %									
parameter k	1	3	5	7	10	15	20	30	40	50
Pulse length ($\boldsymbol{L}$)	84.3	83.3	85.2	82.4	81.5	82.4	79.6	78.7	80.6	78.7
Pulse frequency ($\boldsymbol{F}$)	68.5	69.4	69.4	68.5	66.7	67.6	69.4	68.5	72.2	72.2
Temporal structure($\boldsymbol{T}$)	26.9	25.9	25.9	25.9	27.8	29.6	28.7	30.6	28.7	30.6
$\boldsymbol{LF}$	45.4	47.2	45.4	42.6	43.5	44.4	46.3	49.1	49.1	51.9
$\boldsymbol{LT}$	17.6	14.8	14.8	15.7	19.4	19.4	23.1	23.1	25.9	26.9
$\boldsymbol{FT}$	13.0	11.1	13.0	13.9	14.8	14.8	15.7	15.7	19.4	18.5
$\boldsymbol{LFT}$	8.3	6.5	7.4	7.4	9.3	10.2	11.1	10.2	11.1	13.0

The best classification result based on a single feature is achieved with the temporal structure of pulses ($\boldsymbol{T}$) (error rate ≈ 28 %). This result is significantly improved using additional features from the frequency and time domain. We use the pulse frequency ($\boldsymbol{F}$) and the pulse length (***L***). Whereas the classifier performances based on the single features ***F*** or ***L*** are really bad with error rates of approximately 65-85 %, the combination of all three features (***FLT***) performs significantly better than the best single feature ***T***. For this feature combination (***FLT***) the error rates are in the range of $6 - 10$ % depending on k. This is a very good result for such a multi-class pattern recognition problem with 28 different categories. The confusion matrices of the fuzzy-k-nearest-neighbour classifiers based on the single local features (column 1: pulse length (***L***), column 2: pulse frequency (***F***) and column 3: temporal structure of pulses (***T***)) and their combination (***FLT***) given in Figure 10 show the classification results in more detail.

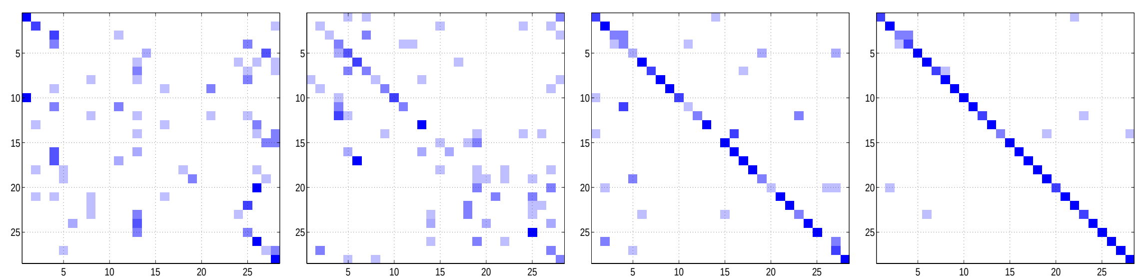


Figure 10: Mean of the confusion matrices of the cross validation runs. ***L***) pulse length, ***F***) pulse frequency, ***T***) temporal structure of the pulses, local features combined through data fusion.

Here it can be observed that feature ***T*** is the best local feature with a small number of positive confusion frequencies.

Comparing the classification results of the global and the local features (see Table II and III) points out that the local features are better suited for the classification, particularly the combination of all three local features. Combining the decisions of the local and global features leads to the classification results depicted in Table IV.

Table IV: The classification results for the combined local and global classifier outputs utilizing average fusion are shown for different k in the range of $k = 1$ to $k = 50$. In the first column (see L↓) are the classification results for the local features and in the first row (see G→) are the classification results for the global features.

G & L	Classification error in %										
Par. k	L↓	1	3	5	7	10	15	20	30	40	50
G→		*35.1*	*34.3*	*43.5*	*43.5*	*41.7*	*45.4*	*44.4*	*41.7*	*40.7*	*40.7*
1	*8.3*	28.7	12.0	9.3	9.3	8.3	8.3	7.4	7.4	8.3	8.3
3	*6.5*	29.6	11.1	9.3	8.3	8.3	6.5	6.5	6.5	6.5	6.5
5	*7.4*	29.6	12.0	9.3	7.4	7.4	7.4	6.5	6.5	6.5	6.5
7	*7.4*	30.6	13.9	9.3	7.4	7.4	7.4	6.5	6.5	6.5	6.5
10	*9.3*	30.6	15.7	9.3	7.4	7.4	7.4	7.4	6.5	6.5	6.5
15	*10.2*	30.6	16.7	9.3	8.3	7.4	8.3	7.4	6.5	8.3	8.3
20	*11.1*	31.5	15.7	10.2	9.3	7.4	8.3	7.4	7.4	7.4	7.4
30	*10.2*	31.5	15.7	12.0	10.2	10.2	8.3	8.3	7.4	7.4	7.4
40	*11.1*	31.5	19.4	13.9	10.2	11.1	9.3	9.3	7.4	7.4	7.4
50	*13.0*	31.5	20.4	13.9	12.0	11.1	12.0	11.1	10.2	9.3	9.3

Table IV also contains the classification rates for the global feature combination (***DCE***) and the local feature combination (***FLT***) from Table II and III. For the local classifiers and the final classification the best error rate is 6.5 %, so the performance of the best local classifier can not be improved using additional information from the global features.
The classification error of all fuzzy-k-nearest-neighbour classifiers (based on global features, local features, and the combination of global and local features) dependends of the parameter k. With this parameter k the fuzzyness of the classifier output is determined. In Table IV it can be observed that it can be useful to combine the fuzzy-k-nearest neighbour decisions of the global and local features with different values of k. The best classification results for the combination of the global and local features are achieved with a lower fuzzyness level for the more accurate classifier (local features, $k = 3-10$) and a higher fuzzyness level for the weaker classifier (global features, $k \geq 10$). This is due to the fact that the correct class of the classifiers of the local features is found in 98 % within the largest four soft labels, but the correct class of the classifiers of the global features is found in 98 %

within the largest 17 soft labels.

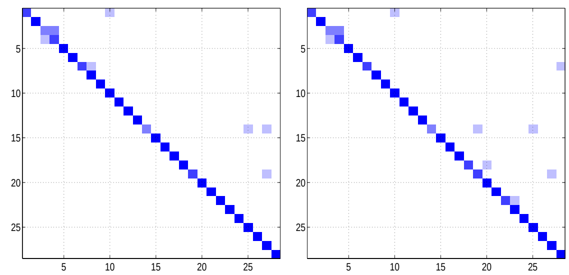


Figure 12: Mean of the confusion matrices of the cross validation rund of the combined local and global a) average fusion, b) symmetrical probabilistic fusion

In Figure 12 the confusion matrices for the combination of local and global decisions built by average and symmetrical probabilistic fusion is shown. Here it can be observed that the errors appear only for a few species.

It must be emphasized that typically a certain feature is not important for all species. For example feature $\boldsymbol{T}$ is more important for species producing *quasi-deterministically oscillating time series* (see Figure 5) [BD87]. Here the distances between the pulses Δ_j create small regions of prototypes in the feature space of the $d-$tuple codes. On the other hand some species produce *stochastically oscillating time series*, where the future evolution can not be determined. For crickets the oscillation of the pulse distances is often quasi-deterministic because the condition for strong periodicity $x(t+T) = x(t)$, where T is the period of the oscillation, is never completely fulfilled.

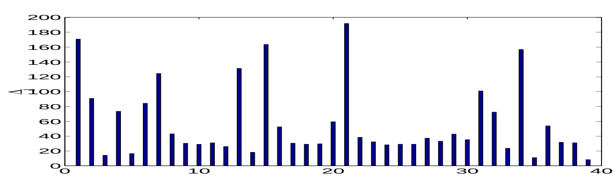


Figure 14: The space between the pulses Δ_j of a $2.5sec$ sound signal of the species *Zvenella geniculata*

In contrast the pulses Δ_j of the *Zvenella geniculata* creates straggled prototypes all over the feature space of the d-tuple codes (see Figure 14), leading to misclassifications.

VII. CONCLUSION

In this study cricket songs are classified utilizing different types of features extracted from the time signal, classification fusion and a fuzzy k-nearest-neighbour classifier. Fuzzy classification techniques are deemed to be a viable extension of classical ones towards handling nonstochastic uncertainty involved in the classification process [Kun96], especially if multiple classification results are combined into one.

The classifiers based on single features are not suited for this recognition task because their error rates are in the range of 25 % to 85 % (for the global features: 45 % to 65 %, see Table II; for the local features: 25 % to 85 %, see Table III). The combination of multiple features leads to a high performance of the classifier system. In particular the fusion of the local features is very promising, leading to an error rate of 6.5 % (see Table III), which is a good result for such a multi-class pattern recognition problem. The best classification results with local features are achieved taking $k = 3$, 5 or 7 nearest neighbours into account. Fusion of global features also leads to a better classification performance, but the error rates are still approximately 35 % $-$ 40 %.

A Combinination of global and local features through classifier fusion can not improve the classification results of the local features (6,5 % error rate). In Figure 12 it can be observed that the misclassifications appear between a small set of species and that sound patterns of 22 species can be categorized without any errors. Furthermore, we found that the correct class is among the two largest soft labels in more than 98 % of all cases.

The study also shows that cricket songs can be classified reliably by artificial neural networks. Basically, they had to solve the same task as cricket females with their "real" neuronal substrate. The results presented here might serve as a heuristic tool for new experiments to understand cricket phonotaxis and neural processing, especially for tropical species. Up to now, neurophysiological experiments have been limited to *Gryllus* spp., which inhabit acoustically less complex biotopes, with much lesser cricket diversity. Though it might be naive to assume that there is any such thing as a neuronal correlate of the mechanisms described here, our results provide important hints for crucial experiments. Parzen density emerged as an important feature which has been overlooked by traditional neurobiologists [Gla96]. Finally, and independent from these questions of fundamental research, the classifiers described here could be implemented into *Rapid Assessment protocols*, to detect, classify and monitor tropical cricket diversity.

ACKNOWLEDGEMENT

DORSA forms part of the Entomological Data Information System (EDIS) and is funded by the German Ministry of Science and Education (BMBF). We are greatful to Sigfrid Ingrisch (ZFMK Bonn, Germany) and Frank Nischk, for providing their sound recordings, suggestions and discussions.

REFERENCES

[BD87] P. J. Browell and R. A. Davis. *Time series: Theory and Methods.* Springer, New York, 1987.

[Bis95] C. M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, New York, 1995.

[BJ92] R. Beale and T. Jackson. *Neuronal Computing.* IOP Publishing Inc., University of York, 1992.

[CFS98] E. D. Chesmore, O. P. Femminella, and M. D. Swarbrick. Automated analysis of insect sounds using time-encoded signals - a new method for species identification. *CAB International*, pages 273–287, 1998.

[EE94] M. R. Evans and J. A. Evans. A computer-based technique for the quantitative analysis of animal sounds. *International Journal of Animal Sound Recording*, 5:281–290, 1994.

[Fuk90] K. Fukunaga. *Introduction to Statistical Pattern Recognition.* Academic Press, New York, 1990.

[Gla96] M. I. Glavinovic. Comparison of parzen density and frequency histogram as estimators of probability density functions. *European Journal of Physiology*, 433(1):174–179, 1996.

[HKA95] D. L. Hawksworth and M. T. Kalin-Arroyo. Magnitude and distribution of biodiversity. *Global Biodiversity Assessment*, United Nations Environment Programme - Cambridge University Press:107–191, 1995.

[HML89] F. Huber, T. E. Moore, and W. Loher. *Cricket Behavior and Neurobiology.* Cornell University Press, Ithaca, NY, 1989.

[IK98] S. Ingrisch and G. Koehler. *Die Heuschrecken Mitteleuropas.* Westarp Wissenschaften, Magdeburg, 1998.

[Ing97] S. Ingrisch. Taxonomy, stridulation and development of podoscirtinae from thailand. *Senckenbergiana biologica*, 77:47–75, 1997.

[KB98] L. I. Kuncheva and J. C. Bezdek. An integrated framework for generalized nearest prototype classifier design. *Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(5):437–457, 1998.

[KM98] J. A. Kogan and D. Margoliash. Automated recognition of bird song elements from continous recordings using dynamic time warping and hidden markov models: A comparative study. *Journal of the Acoustical Society of America*, 103(4):2185–2196, 1998.

[Kun96] L. I. Kuncheva. On the equivalence between fuzzy and statistical classifiers. *Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 4(3):245–253, 1996.

[Kun00] L. I. Kuncheva. Clustering and selection model for classifier combination. In *4th International Conference on Knowledge-Based Intelligent Engineering Systems*, 2000.

[MMR98] S. O. Murray, E. Mercado, and H. L. Roitblat. The neural network classification of false killer whale *(pseudorca crassidens)* vocalizations. *Journal of the Acoustical Society of America*, 104(6):3626–3633, 1998.

[Nis99] F. Nischk. *Die Grillengesellschaften zweier neotropischer Waldökosysteme in Ecuador*. PhD thesis, University of Köln in Germany, 1999. Memorandum UCB/ERL–M89/29.

[Ott92] D. Otte. Evolution of cricket songs. *Journal of Orthopteran Research*, I:25–44, 1992.

[Par62] E. Parzen. On the estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33:1065–1076, 1962.

[Pea88] J. Pearl. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann Publishers, San Francisco, 1988.

[PG90] T. Poggio and F. Girosi. Networks for approximation and learning. *Proceedings of the IEEE*, 78(9):1481–1497, 1990.

[Rie98] K. Riede. Acoustic monitoring of orthoptera and its potential for conversation. *Journal of Insect Conservation*, 2:217–223, 1998.

[RJ93] N. Rabiner and L. Y. Juang. *Fundamentals of Speech Recognition.* PTR Prentice-Hall, New Jersey, 1993.

[RS97] L. R. Rabiner and S. E. Sambur. An algorithm for determining the endpoints of isolated utterances. *Bell Syst. Technical Journal*, 54(2):297–315, 1997.

[Rus88] G. Ruske. *Automatische Spracherkennung.* Oldenbourg Verlag, München, 1988.

[Sin98] S. Singh. 2D spiral pattern recognition with possibilistic measures. *Pattern Recognition Letters*, 19(2):141–147, 1998.

[SKP01] F. Schwenker, H. A. Kestler, and G. Palm. Three phase learning for radial basis function networks. *Neural Networks*, 14:439–458, 2001.

[TB96] S. Thrun and A. Brücken. Integrating grid-based and topological maps for mobile robot navigation. *National Conference on Artificial Intelligence Portland, Oregon*, 1996.

[VHH98] P. Vary, U. Heute, and W. Hess. *Digitale Sprachsignalverarbeitung.* B. G. Teubner Stuttgart, 1998.