Classification of Bird Song Syllables Using Wigner-Ville Ambiguity Function Cross-Terms

Maria Sandsten
Mathematical Statistics,
Centre for Mathematical Sciences,
Lund University, Sweden
Email: sandsten@maths.lth.se

Johan Brynolfsson
Mathematical Statistics,
Centre for Mathematical Sciences,
Lund University, Sweden
Email: johanb@maths.lth.se

***Abstract*—A novel feature extraction method for low-dimensional signal representation is presented. The features are useful for classification of non-stationary multi-component signals with stochastic variation in amplitudes and time-frequency locations. Using a penalty function to suppress the Wigner-Ville ambiguity function auto-terms, the proposed feature set is based on the cross-term doppler- and lag profiles. The investigation considers classification where strong similar components appear in all signals and where the differences between classes are related to weaker components. The approach is evaluated and compared with established methods for simulated data and bird song syllables of the great reed warbler. The results show that the novel feature extraction method gives a better classification than established methods used in bird song analysis.**

I. Introduction

In biology, bird song analysis has been a large field for several decades and methods based on spectrograms (sonograms) have been considered well suited for the comparison of bird sounds. Characterizing patterns of songs from different bird species are often sufficiently distinct, so that rather straightforward features, e.g., time duration and frequency bandwidth or cross-correlation of spectrograms (SPCC), often yield satisfactory results, [1], [2]. Somewhat more sophisticated is song analysis by means of e.g., pitch tracking [3], dynamic time warping [4], mel-frequency cepstral coefficients (MFCC) [5], and hidden Markov models [6].

The other main question in bird song research is the automatic clustering of *within-species* syllables, e.g. for computing repertoire size. This task often constitutes a much more involved problem, and requires sufficiently sophisticated methods able to not only capture subtle characteristic details within a song but also to compare them with each other, [7]. More simplistic methods, which may be suitable for species differentiation, smooth out the differences that should be detected and will fail in the within-species analysis. The great reed warbler (GRW) is one example of a species with songs of high complexity. Song analysis for the GRW has so far mainly been conducted manually, by listening and visually studying the syllable sonograms, [8], [9]. One of few successful attempts to automatically cluster the syllables of the GRW song has recently been made in [10], [11]. The features used for clustering are the first pair of singular vectors of the ambiguity function corresponding to a multitaper spectrogram. Multitaper spectrograms have been used earlier for more robust estimation of features, [3], [7] and in a recent submission, it was shown that the multitaper spectrogram is robust to jitters in the component amplitudes and time-frequency locations, [12].

However, to fully focus on the subtle characteristic details, sophisticated methods that consider the small differences in syllable structures are needed. It has been argued and shown that the Wigner-Ville distribution should be better to use for feature extraction than any traditionally smoothed distribution, e.g., the spectrogram, [13]. The Wigner-Ville distribution contains cross-terms, usually considered to be non-relevant and therefore they should be suppressed. In this submission, we focus on the explicit use of the Wigner-Ville ambiguity function cross-terms to capture and classify the subtle differences in stochastic multi-component signals where strong similar components appear in all signals and where the differences between classes are related to weaker components.

II. The ambiguity penalty kernel

The **Wigner-Ville distribution** (WVD) is defined as

$$W_z(t,f) = \int z(t+\frac{\tau}{2})z^*(t-\frac{\tau}{2})e^{-i2\pi f\tau}d\tau, \qquad (1)$$

where $z(t)$ is the analytic correspondence to a real-valued signal, obtained using the Hilbert transform. All integrals in this paper range from $-\infty$ to ∞. We study the two-component signal $z(t) = z_1(t) + z_2(t)$ for which the WVD is,

$$W_z(t,f) = W_{z_1}(t,f) + W_{z_2}(t,f) + 2\Re[W_{z_1,z_2}(t,f)], \quad (2)$$

where $W_{z_1}(t,f)$ and $W_{z_2}(t,f)$, called **auto-terms**, are the WVDs of $z_1(t)$ and $z_2(t)$ respectively and $2\Re[W_{z_1,z_2}(t,f)]$ is referred to as the **cross-term**, where $\Re$ represents the real part. The **ambiguity function** (AF) is defined as

$$A_z(\nu,\tau) = \int z(t+\frac{\tau}{2})z^*(t-\frac{\tau}{2})e^{-i2\pi\nu t}dt, \qquad (3)$$

where signal auto-terms always will be located at the centre, independently of where they are located in the time-frequency plane, and the cross-terms will always be located away from the centre. The natural approach is to keep the terms located at the centre and suppress the components away from the centre

ISBN 978-0-9928626-7-1 © EURASIP 2017

using an **ambiguity kernel**. However, to differ details when signal components are of very different amplitudes, it could be more beneficial to focus on the cross-terms. To illustrate, we study the following simple example with a two-component sinusoidal signal,

$$z(t) = z_1(t) + z_2(t) = c_1 e^{-i2\pi f_1 t} + c_2 e^{-i2\pi f_2 t}, \quad (4)$$

where the absolute value of the resulting AF is

$$\begin{aligned} |A_z(\nu,\tau)| &= (c_1^2 + c_2^2)\delta(\nu) + \ldots \\ & \quad c_1 c_2(\delta(\nu + (f_1 - f_2)) + \delta(\nu - (f_1 - f_2)). \end{aligned}$$

The first term with magnitude $c_1^2 + c_2^2$ corresponds to the auto-terms and the remaining two terms with magnitude $c_1 c_2$ correspond to the cross-term. Consider the case when one component is much weaker than the other, $c_2 << c_1$. The auto-term magnitude, $c_1^2 + c_2^2 \approx c_1^2$, i.e., the existence of a weak amplitude component will be hidden by the large-amplitude component, where on the other hand the cross-term magnitudes $c_1 c_2$ are using c_1 as an amplification. If the auto-terms are suppressed, the cross-terms will clearly show the existence of the weak component as well as its frequency location in relation to the large-amplitude component.

In this paper, the auto-term area is defined from the ambiguity function contour of a single Gaussian function $w(t) = (\beta/\pi)^{1/4} e^{-\frac{\beta}{2}t^2}$, where the corresponding component length N_β^P is the number of samples defining the Gaussian function above $\mu \cdot (\beta/\pi)^{1/4}$. The corresponding AF is

$$A_w(\nu,\tau) = e^{-(\frac{\beta\tau^2}{4} + \frac{\pi^2\nu^2}{\beta})}. \quad (5)$$

A **ambiguity penalty (AP) kernel** is defined as

$$\phi^P(\nu,\tau) = 0 \quad \text{for } A_w(\nu,\tau) > \mu, \quad (6)$$

and one for all other values. In this paper $\mu = 0.01$. The resulting absolute value of the ambiguity function using the kernel $\phi^P(\nu,\tau)$ is

$$|A_z^P(\nu,\tau)| = |A_z(\nu,\tau) \cdot \phi^P(\nu,\tau)|. \quad (7)$$

The resulting cross-term normalized doppler- and lag profiles are defined as

$$\begin{aligned} M_D(\nu) &= \frac{1}{E_A} \int |A_z^P(\nu,\tau)| d\tau, \\ M_L(\tau) &= \frac{1}{E_A} \int |A_z^P(\nu,\tau)| d\nu, \end{aligned} \quad (8)$$

where $E_A = \int\int |AP_z(\nu,\tau)| d\tau d\nu$. As measure of similarity of the two syllables, $s1$ and $s2$, the inner product in Euclidean space defined by,

$$d_D(s1,s2) = \langle M_D^{s1}, M_D^{s2} \rangle = \int M_D^{s1}(\nu) M_D^{s2}(\nu) d\nu, \quad (9)$$

is used for the doppler profile vectors. The lag profile vectors are combined similarly and the resulting two measures are then averaged, i.e.,

$$d(s1,s2) = (d_D(s1,s2) + d_L(s1,s2))/2, \quad (10)$$

so the best possible similarity is $d(\cdot,\cdot) = 1$, where orthogonal vectors indicate difference and $d(\cdot,\cdot) = 0$ is the smallest possible value.

III. Example

A synthetic bird song syllable model is proposed as

$$x(n) = \sum_{j=1}^{J} A_j \cos(2\pi F_j n + \phi) \cdot w_j(n - T_j), \quad (11)$$

where $n = 0 \ldots N-1$, $\phi \in R(0, 2\pi)$. The function $w_j(n) = e^{-\frac{\alpha_j}{2} n^2}$ is a Gaussian window with α_j chosen such that N_j^g values are above μ. To exemplify and further explain the advantages in the case of large amplitude differences between components, three different syllables are simulated, with $N = 800$, all $N_j^g = 128$ and the other parameters according to Table I. The syllables $s1_{C1}$ and $s2_{C1}$ belong to class 1 where $s1_{C2}$ belongs to class 2.

Par.	$s1_{C1}$	$s2_{C1}$	$s1_{C2}$
$A_1/T_1/F_1$	1 / 200 / 0.1	1 / 200 / 0.1	1 / 200 / 0.1
$A_2/T_2/F_2$	0.1 / 500 / 0.1	0.1 / 500 / 0.1	0.1 / 200 / 0.2

TABLE I
THE PARAMETERS OF THE THREE SYLLABLES IN THE EXAMPLE.

The results of the similarity measure in Eq. (10), using the proposed penalty kernel from Eq. (6) with $N_\beta^P = N_j^g = 128$, are presented in Table II (WIGAP). To show the actual gain of the penalty function, the corresponding measures, with $A_z^P(\nu,\tau)$ replaced by just the Wigner-Ville ambiguity function $A_z(\nu,\tau)$ in Eq. (8), are computed, (WIGA). The WIGAP shows a clear difference between the in-class measure (1.0) and the between-class measures (0.038 and 0.041) where WIGA indicates the same similarity between all syllables. The reason is obvious, the high-amplitude components are at the same time- and frequency location and the power of the small component is about 1% of the large component, which will not be visible as a difference in the time- and frequency profiles.

Meas.	$d(s1_{C1}, s2_{C1})$	$d(s1_{C1}, s1_{C2})$	$d(s2_{C1}, s1_{C2})$
Desired	1.0	0	0
WIGAP	1.0	0.038	0.041
WIGA	1.0	0.99	0.99

TABLE II
THE SIMILARITY MEASURES OF DIFFERENT COMBINATIONS OF THE THREE SYLLABLES USING DIFFERENT METHODS.

IV. Evaluation

In the evaluation, the signals are defined as in Eq. (11) but with stochastic component parameters according to Table III, for the two different classes. The amplitude, frequency- and time locations are stochastic variables with Gaussian distributions, $A_j \in N(A_j^0, \sigma_{A_j})$, $F_j \in N(F_j^0, \sigma_{F_j})$ and

ISBN 978-0-9928626-7-1 © EURASIP 2017

$T_j \in N(T_j^0, \sigma_{T_j})$, where a number of 100 realizations is generated in each class. The realizations are also disturbed by white zero-mean Gaussian noise with variance σ_N^2. The signal-to-noise ratio (SNR) is defined by

$$\text{SNR} = 10\log_{10}\frac{P_a}{\sigma_N^2}, \tag{12}$$

where P_a is the average of the total energy of all 100 realizations. An example for SNR=12 dB is shown in Fig. 1a), where one could note that the specified SNR-measure is not fair for the low-amplitude components, which are the ones carrying the differences between the two classes. The local SNR, measured just for the low-amplitude component (red color), is in this case 0 dB. The WIGAP and WIGA applied in the previous example are also used here. For comparison, the MFCC-method with 8 cepstral coefficients, a 128 sample Hamming window and 90% overlap between frames, is investigated [14], as well as the SPCC method with 128 sample Hanning window spectrograms. Additionally, the multitaper SVD-based method using the second pair of singular vectors, [12], with 8 multitapers and the first Gaussian window of length 100 samples (SVDMT), is evaluated.

Class 1	A_j^0, σ_{A_j}	F_j^0, σ_{F_j}	T_j^0, σ_{T_j}
$j=1$	$1, \sigma$	$0.1, 0.5\sigma$	$200, 800\sigma$
$j=2$	$0.1, \sigma$	$0.1, 0.5\sigma$	$500, 800\sigma$
Class 2	A_j^0, σ_{A_j}	F_j^0, σ_{F_j}	T_j^0, σ_{T_j}
$j=1$	$1, \sigma$	$0.1, 0.5\sigma$	$200, 800\sigma$
$j=2$	$0.1, \sigma$	$0.2, 0.5\sigma$	$200, 800\sigma$

TABLE III
THE PARAMETERS OF THE TWO CLASSES IN THE EVALUATION.

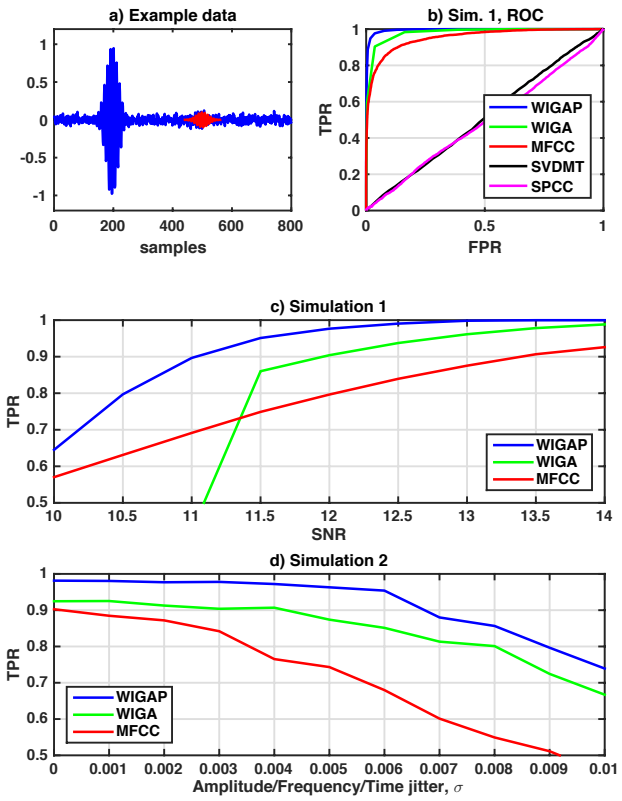

Fig. 1. a) An example signal for SNR=12 dB, also illustrating the low local SNR=0 dB of the low-amplitude component (red color); b) The ROC curves for SNR=12 dB and all the different methods; c) Simulation 1 with $\sigma = 0.004$ and varying SNR: True positive rates for FPR 5%; d) Simulation 2 with SNR=12 dB and varying jitters σ in model parameters: True positive rates for FPR 5%.

In the first simulation, the stochastic jitter parameter σ of Table III is 0.004 and the disturbing noise σ_N is varied giving an SNR from 10 to 14 dB. The results of all methods for SNR=12 dB are shown as Receiver Operating Characteristics (ROC) in Fig. 1b), where it is clear that SVDMT and SPCC (the diagonal lines) both fail as they perform similarly to random classification. The results will be the same for all parameter values, which exclude them for the further analysis. The WIGAP, WIGA and MFCC are the only methods that are able to differ between the classes. In Fig. 1c), the true positive rates (TPR) accepting a false positive rate (FPR) of 5%, are depicted for the different SNRs. The results show that WIGAP gives well above 95% TPR down to 12 dB, where WIGA and MFCC both perform much worse. The WIGA does not work at all for lower SNRs than 11.5 dB.

In the second simulation, the SNR is fixed to 12 dB and the parameter σ is varied in the evaluations, increasing the jitter in amplitudes and time- and frequency locations of the components, according to Table III. The results are presented in Fig. 1d), where the different TPR, accepting FPR 5%, for WIGAP, WIGA and MFCC are shown. The WIGAP gives a TPR of 95% up to $\sigma = 0.006$, where WIGA as well as MFCC are below 85% and 70% respectively.

V. REAL DATA EVALUATION

The methods are evaluated on a small data set of three hand-sorted classes of syllables recorded from one individual of the GRW, depicted in Figs. 2a-c). The sample frequency is 11 kHz and the syllables of a class are time-aligned using ordinary time-based correlation. A spectrogram from an example of each class, marked with red color in the Figs. 2a-c), is shown in Figs. 2d-f). The frequency contents are more or less the same as well as the time support, although the pitch frequencies of class 1 are somewhat higher than in the other two classes. There are clear differences in the distribution of the weaker components between classes.

In Fig. 3, the upper part of the ROC-curves (note the scale of the y-axis) are depicted for a pairwise analysis of the three classes and all methods. The parameters of the methods are changed for a better fit to the real data, i.e., in accordance with the large amplitude component lengths of 20-30 ms (220-330 samples), $N_\beta^P = 256$ for WIGAP, a 256 sample Hanning window for SPCC, a 256 sample Hamming window of MFCC and finally 8 multitapers with the first Gaussian window of length 200 samples are applied for the SVDMT. In Fig. 3a), the classification between class 1 and 2 is performed very well

ISBN 978-0-9928626-7-1 © EURASIP 2017

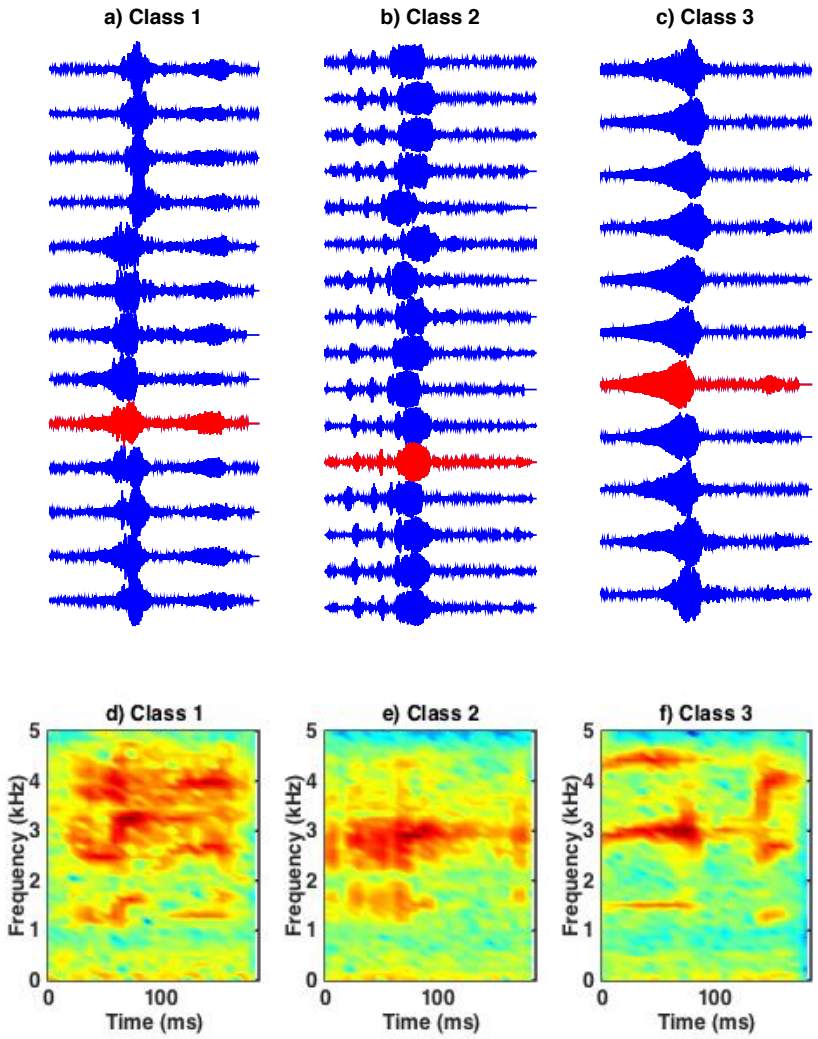

Fig. 2. a-c) The included syllables of the three classes; d-f) spectrogram of one syllable from each class.

for WIGAP, WIGA and MFCC, that all give a TPR of 100%. In Fig. 3b), the classification is more difficult as there are stronger similarities between these two classes. The WIGAP and WIGA perform well, with WIGAP giving 100% TPR and WIGA slightly below. All other methods, including the MFCC, have a bad performance. The last classification is between class 2 and 3 and the results are shown in Fig. 3c), where WIGAP gives the highest TPR followed by WIGA and MFCC.

VI. Conclusions

A novel penalty kernel is suggested, with the aim to eliminate the auto-terms and keep the cross-terms of the Wigner-Ville ambiguity function. The resulting doppler frequency and lag profiles are used as features for classification of signals with large differences in component amplitudes. The results show that the proposed kernel contributes to a classification that is robust to additive disturbing noise as well as to stochastic variation in amplitudes and time- and frequency locations. Evaluation for a real data set of syllables from the great reed warbler shows that the classification based on the novel kernel outperforms the established methods, such as SPCC and MFCC.

VII. Acknowledgments

The authors would like to thank the Swedish strategic research programme eSSENCE for funding and the department of Biology Lund University for data collection.

References

[1] E. R. A. Cramer, "Measuring consistency: spectrogram cross-correlation versus targeted acoustic parameters," *Bioacoustics: The International Journal of Animal Sound and its recording*, vol. 22, no. 3, pp. 247–257, 2012.

[2] S. Keen, J. C. Ross, E. T. Griffiths, M. Lanzone, and A. Farnsworth, "A comparison of similarity-based approaches in the classification of flight calls of four species of north american wood-warblers (parulidae)," *Ecological Informatics*, vol. 21, pp. 25–33, 2014.

ISBN 978-0-9928626-7-1 © EURASIP 2017

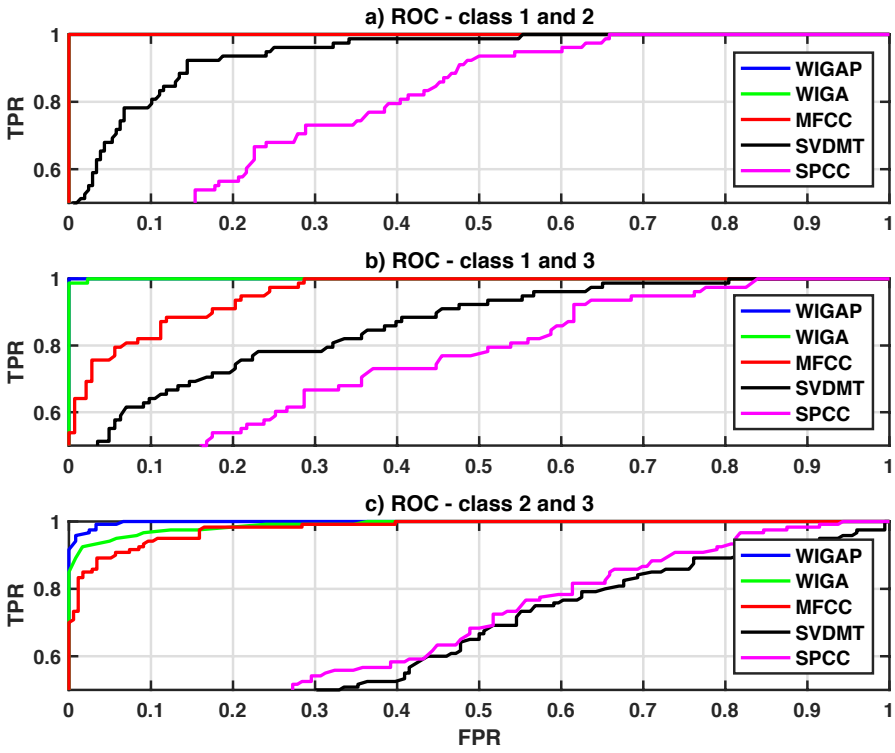

Fig. 3. The ROC curves (upper half) for pairwise classification.

[3] C. D. Meliza, S. C. Keen, and D. R. Rubenstein, "Pitch- and spectral-based dynamic time warping methods for comparing field recordings of harmonic avian vocalizations," *J. Acoust. Soc. Am.*, vol. 134, no. 2, pp. 1407–1415, 2013.

[4] L.N. Tan, A. Alwan, G. Kossan, M.L. Cody, and C.E. Taylor, "Dynamic time warping and sparse representation classification for birdsong phrase classification using limited training data.," *Journal of the Acoustical Society of America*, vol. 137, no. 3, pp. 1069 – 1080, 2015.

[5] P. Somervuo, A. Härmä, and S. Fagerlund, "Parametric representations of bird sounds for automatic species recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 6, pp. 2252–2263, 2006.

[6] P. Jancovic, M. Kokuer, M. Zakeri, and M. Russell, "Bird species recognition using HMM-based unsupervised modelling of individual syllables with incorporated duration modelling.," in *ICASSP- Proceedings*, 2016, vol. 2016-May, pp. 559–563.

[7] O. Tchernichovski, F. Nottebohm, C. E. Ho, B. Pesaran, and P. P. Mitra, "A procedure for an automated measurement of song similarity," *Animal Behaviour*, vol. 59, no. 6, pp. 1167–1176, 2000.

[8] E. Węgrzyn and K. Leniowski, "Syllable sharing and changes in syllable repertoire size and composition within and between years in the great reed warbler, *acrocephalus arundinaceus*," *J. Ornithol.*, vol. 151, pp. 255–267, 2010, DOI 10.1007/s10336-009-0451-x.

[9] D. Hasselquist, S. Bensch, and T. von Schantz, "Correlation between male song repertoire, extra-pair paternity and offspring survival in the great reed warbler," *Nature*, vol. 381, pp. 229–232, 1996.

[10] M. Große Ruse, D. Hasselquist, B. Hansson, M. Tarka, and M. Sandsten, "Automated analysis of song structure in complex bird songs," *Animal Behaviour*, vol. 112, pp. 39–51, 2016.

[11] M. Sandsten, M. Große Ruse, and M. Jönsson, "Robust feature representation for classification of bird song syllables," *EURASIP Journal on Advances in Signal Processing*, 2016, DOI:10.1186/s13634-016-0365-8.

[12] M. Hansson-Sandsten, "Classification of bird song syllables using singular vectors of the multitaper spectrogram," in *European Signal Processing Conference (EUSIPCO)*, Nice, France, 2015.

[13] B. W. Gillespie and L. E. Atlas, "Optimizing time-frequency kernels for classification," *IEEE Trans. on Signal Processing*, vol. 49, no. 3, pp. 485–496, 2001.

[14] M. Slaney, "Auditory toolbox: Version 2," Technical Report No. 1998-010, https://engineering.purdue.edu/malcolm/interval/1998-010/

ISBN 978-0-9928626-7-1 © EURASIP 2017