Sandsten *et al. EURASIP Journal on Advances in Signal Processing* (2016) 2016:68
DOI 10.1186/s13634-016-0365-8

EURASIP Journal on Advances in Signal Processing

RESEARCH Open Access

Robust feature representation for classification of bird song syllables

Maria Sandsten[1*], Mareile Große Ruse[2] and Martin Jönsson[2]

Abstract

A novel feature set for low-dimensional signal representation, designed for classification or clustering of non-stationary signals with complex variation in time and frequency, is presented. The feature representation of a signal is given by the first left and right singular vectors of its ambiguity spectrum matrix. If the ambiguity matrix is of low rank, most signal information in time direction is captured by the first right singular vector while the signal's key frequency information is encoded by the first left singular vector. The resemblance of two signals is investigated by means of a suitable similarity assessment of the signals' respective singular vector pair. Application of multitapers for the calculation of the ambiguity spectrum gives an increased robustness to jitter and background noise and a consequent improvement in performance, as compared to estimation based on the ordinary single Hanning window spectrogram. The suggested feature-based signal compression is applied to a syllable-based analysis of a song from the bird species Great Reed Warbler and evaluated by comparison to manual auditive and/or visual signal classification. The results show that the proposed approach outperforms well-known approaches based on mel-frequency cepstral coefficients and spectrogram cross-correlation.

Keywords: Time-frequency analysis, Ambiguity spectrum, Singular value decomposition, Multitaper, Bird song

1 Introduction

In biology, bird song analysis has been a large field for several decades, and for many years, methods based on spectrograms (sonograms) have been considered well-suited for the comparison of bird sounds. Generally, song analysis tools are especially challenged when recordings have been conducted on birds under natural outdoor conditions. In these environments, disturbing background noises, such as wind and interference from other birds, is typically present and often distorts the recorded signal substantially. The extent to which such background noise effectively impairs the analysis depends on the type/structure of the underlying signal and on the particular research question. Essentially two principal topics have been considered in literature in the context of classification or clustering of bird song units (e.g., syllables). The hitherto most common research aims at the song-based identification of bird species. Characterizing patterns of songs from *different bird species* are often sufficiently distinct, so that rather straightforward features such as time and frequency moments, time duration, and frequency bandwidth often yield satisfactory results. Somewhat more sophisticated is song analysis by means of pairwise cross-correlation of spectrograms (SPCC), [1, 2] or dynamic time warping (DTW), [3–5]. Besides, methods that have become popular in speech analysis, such as approaches based on pitch frequency or mel-frequency cepstrum coefficients (MFCC), have been successfully applied to bird species classification, [6]. More recently, bird sounds analysis of especially noisy signals has been approached using wavelets, [7].

The other main question guiding bird-song research is the *within-species* classification and clustering. This task often constitutes a much more involved problem, especially when the songs of the species under consideration have a complex structure. Such problems thus require sufficiently sophisticated methods that are able to not only capture subtle characteristic details within a song, but also to compare them with each other. More simplistic methods, which may be well-tailored for species identification, smooth out the differences that should be detected and will fail in the within-species analysis. The Great Reed

*Correspondence: sandsten@maths.lth.se
[1] Lund university, Mathematical Statistics, Centre for Mathematical Sciences, Lund, Sweden
Full list of author information is available at the end of the article

Warbler (GRW) is one example of a species with songs of pronounced profound complexity. However, due to the lack of sufficiently sophisticated methods, song analysis for the GRW has so far mainly been conducted manually, by listening and visually studying the syllable sonograms, [5, 8, 9].

Bird song analysis is only one of various several applications of time-frequency (TF) analysis of non-stationary signals and a significant number of approaches in this important field has been suggested. Since its introduction to bird song analysis in the 50 s, the sonogram (or TF spectrum) has become one of the most established tools in context of bird song investigation, and computationally efficient and robust algorithms for spectrogram-based TF analysis can be found using, e.g., multitapers (MTs). Originally, MTs were introduced by Thomson, [10], who proposed the discrete Prolate Spheroidal Sequences for estimation of a low-variance spectrum with pre-specified resolution. Nowadays, the noise-robust Thomson MTs are well-established in the context of stationary spectrum analysis and have found various application areas. Recently, the *Hermite* MTs, [11], have gained popularity, especially in applications with particular interest in estimation of the TF spectrum of non-stationary signals, [12–16]. As the MT spectrogram is known to reduce variation in amplitude and to limit resolution in time and frequency, it is tailored for the analysis of multi-component signals with jitter, or variance, in both location of the components and their amplitudes, [17]. In contrast to TF distributions, that aim at optimal resolution of signal components and cross-term suppression, [18], MT spectrograms are more suitable for the type of data considered in this paper, as MTs are expected to smooth out small differences in time and frequency locations and therefore lower the in-class variance.

Extracting features from the TF spectrum for classification or clustering is a non-trivial task, and several approaches have been proposed in literature. In different application areas, there has recently been an increased interest in decomposition methods, such as approaches related to principal component analysis (PCA), independent component analysis (ICA), singular value decomposition (SVD), and non-negative matrix factorization (NMF). In [18], the authors achieve increased noise-insensitivity by combining image processing techniques with wavelets and SVD. Barry et al., [19] improved classification performance of event-related signals by the application of PCA to TF spectra of the electroencephalogram data. The SVD of TF spectra was also used to classify multi-component bird song syllables in [17], and multi-component frequency modulated (FM) signals in [20]. Approaches employing these techniques are promising for two key reasons. On the one hand, they create a decoupling of the time and frequency domains and therefore facilitate the separate inspection of corresponding features, and on the other hand, they achieve a noise reduction as noise is spread across the collection of all singular vectors and the signal part usually has low rank [19, 20]. The NMF, [21], where a matrix with positive values is decomposed into positive basis functions, has been applied for classification of TF spectrum of audio signals [22, 23]. In [24], it is shown that the NMF decomposition method of a TF spectrum is superior to PCA and ICA for classification of audio data. However, the NMF method is computationally more demanding than, e.g., the SVD technique, as it requires an iterative solution. Different algorithms for better convergence have been proposed, and recently, an approach using the SVD basis functions as initialization of the NMF algorithm was suggested, [25].

A less intuitive, but for certain topics—particularly in bird song analysis—highly suitable tool for signal representation and the ground for feature extraction is the so-called ambiguity spectrum (AS) [26]. A characteristic for this two-dimensional Fourier transform of the TF spectrum is its invariance to time and frequency shifts of the signal. More specifically, the absolute values of the AS of a signal and its time and frequency shifted version are identical. Therefore, signal analysis based on the AS focuses on differences between time and frequency components rather than on their actual location in the TF plane. Moreover, for many applications—as also for representation of syllables from bird songs—the AS will be a matrix of low rank, and can hence be well-approximated by only a few (e.g., the first pair) of its singular vectors. The AS and its first pair of singular vectors will constitute the essence of our methodology.

As the MT spectrogram is robust to jitters in the amplitudes and locations of a signal's components [17], selecting the first two singular vectors of the MT spectrogram is more intuitive than the AS-based representation. However, as shown in [17], using singular vectors of the spectrogram for classification of multi-component signals requires several singular vector pairs and more advanced algorithms to combine the singular vectors in an appropriate way. To ensure better comparability of ambiguity- and TF-domain analysis, we, however, base our investigations in both cases on only the first pair of singular vectors.

The main contribution of this paper is the introduction of a feature set based on SVD of the AS on the basis of which, e.g., classification and clustering tasks of non-stationary signals can be performed. The latter may be conducted in terms of a similarity measure. The method has recently been applied to clustering of a whole song of the GRW [27], where in this work, a collection of possible similarity measures is presented and their respective performances are evaluated and compared. Additionally, the optimal parameters of the TF methods are found,

and robustness of methods are investigated for additional noise disturbances.

The suggested algorithm consists of four steps. The detection step aims at detecting and subdividing a bird-song strophe into individual syllables ($\sim$ 50–300 ms). In the second step, the syllable-specific ambiguity spectra are estimated, and the corresponding SVDs are calculated in the third step. Each syllable will then be represented by the first two singular vectors of its AS. As the ambiguity matrix for these kinds of signals is typically of low rank, this representation captures the signal's key information, both in time and frequency direction. In the fourth step, the alikeness of two syllables, represented by their respective pair of singular vectors, is assessed by means of a collection of candidate similarity measures, which are evaluated and compared to each other.

The reminder of the article is structured as follows. In the subsequent section, we give a short treatment of the TF representation of a signal along with the quadratic class of smoothed spectra. The ambiguity spectrum, which will play an important role in our methodology, is introduced and its utilization is motivated. We give a first application of some spectral methods on a bird-song signal, the latter being the main object for our analysis. In section 3, we introduce our feature set which is based on the SVD of a signal's AS and provide a few examples. Next, we present two raw similarity measures in section 4 and use them to derive three combined measures, all of which subsequently are to be assessed and compared. In section 5, we describe a two-step method for detection of syllables from a bird song strophe. The data used for our examples is described in section 6, and a baseline truth for the classification is defined. In section 7, we evaluate the suggested similarity measures as well as different approaches for estimation of the AS. Moreover, this section provides a comparison of the proposed approach to other well-known methods. Section 8 contains a major application of our methodology to a larger set of syllables in a more complex classification study. Finally, section 9 concludes.

2 Time-frequency analysis and multitapers

For a non-stationary signal, the instantaneous autocorrelation function (IAF) of a zero-mean signal $x(t)$ is a function of two variables t and τ defined as

$$r_x(t,\tau) = E\left[x\left(t+\frac{\tau}{2}\right)x^H\left(t-\frac{\tau}{2}\right)\right], \tag{1}$$

where $E[*]$ denotes the expectation operator and the superscript H the conjugate-transpose. The Wiener-Khintchine theorem extended to the time-*varying* spectral density and an application of the Fourier transform with respect to the variable τ give the so-called Wigner or TF spectrum, [28], $W_x(t,f) = \mathcal{F}_{\tau\to f} r_x(t,\tau)$. For a given TF kernel $\Phi(t,f)$, we find in the quadratic class (Q) the *smoothed TF spectrum* as the two-dimensional convolution,

$$W_x^Q(t,f) = W_x(t,f) **\Phi(t,f). \tag{2}$$

The AS spectrum is obtained by application of the Fourier transform with respect to the variable t in the IAF, $A_x(\nu,\tau) = \mathcal{F}_{t\to\nu} r_x(t,\tau)$. For a given ambiguity domain kernel $\phi(\nu,\tau)$, one defines the *filtered AS* [28], as

$$A_x^Q(\nu,\tau) = A_x(\nu,\tau)\cdot\phi(\nu,\tau). \tag{3}$$

The relationship between the smoothed TF spectrum, the filtered AS and the IAF is given by

$$W_x^Q(t,f) = \int\int A_x^Q(\nu,\tau)e^{-i2\pi(\tau f-t\nu)}d\tau d\nu \tag{4}$$

$$= \int\int A_x(\nu,\tau)\phi(\nu,\tau)e^{-i2\pi(\tau f-t\nu)}d\tau d\nu \tag{5}$$

$$= \int\int r_x(u,\tau)\rho(t-u,\tau)e^{-i2\pi f\tau}dud\tau, \tag{6}$$

with *time-lag kernel* $\rho(t,\tau) = \mathcal{F}^{-1}_{\nu\to t}\phi(\nu,\tau)$. Using the change of variables, $t = (t_1+t_2)/2$ and $\tau = t_1 - t_2$, Eq. (1) becomes

$$r_x((t_1+t_2)/2, t_1-t_2) = E[\,x(t_1)x^H(t_2)\,], \tag{7}$$

and therefore Eq. (6) can be rewritten as

$$\begin{aligned} W_x^Q(t,f) &= \int\int E[\,x(t_1)x^H(t_2)\,]\,\rho\left(t-\frac{t_1+t_2}{2}, t_1-t_2\right) \\ &\quad\times\ e^{-i2\pi f(t_1-t_2)}dt_1dt_2 \\ &= \int\int E[\,x(t_1)x^H(t_2)\,]\,\rho^{rot}(t-t_1, t-t_2) \\ &\quad\times\ e^{-i2\pi ft_1}e^{i2\pi ft_2}dt_1dt_2, \end{aligned} \tag{8}$$

with

$$\rho^{rot}(t_1,t_2) = \rho\left(\frac{t_1+t_2}{2}, t_1-t_2\right). \tag{9}$$

In general, if the kernel $\rho^{rot}(t_1,t_2)$ satisfies the Hermitian property

$$\rho^{rot}(t_1,t_2) = (\rho^{rot}(t_2,t_1))^H, \tag{10}$$

the solution of the integral equation

$$\int \rho^{rot}(t_1,t_2)q(t_1)dt_1 = \lambda q(t_2), \tag{11}$$

results in eigenvalues λ_k and eigenfunctions $q_k, k\in\mathbb{N}$, which form a complete set. The time-lag kernel can then be expressed as

$$\rho^{rot}(t_1,t_2) = \sum_{k=1}^{\infty}\lambda_k q_k(t_1)q_k^H(t_2). \tag{12}$$

Using the eigenvalues and eigenvectors, Eq. (6) is rewritten as a weighted sum [29],

$$W_x^Q(t,f) = \sum_{k=1}^{\infty} \lambda_k E[|\int x(t_1) e^{-i2\pi f t_1} q_k(t-t_1) dt_1|^2] \quad (13)$$

$$= \sum_{k=1}^{\infty} \lambda_k S_x^{(k)}(t,f) = S_x(t,f), \quad (14)$$

where each $S_x^{(k)}$ is a spectrogram with window function q_k. Thus, the corresponding filtered AS can be calculated as $A_x^Q(\nu,\tau) = \mathcal{F}_{t\to\nu}\mathcal{F}_{f\to\tau} S_x(t,f)$.

Depending on the eigenvalues λ_k, the number of different spectrograms that are averaged could be just a few or an infinite number. In particular, if only finitely many, say K, eigenvalues are non-zero, one has

$$S_x(t,f) = \sum_{k=1}^{K} \lambda_k S_x^{(k)}(t,f), \quad (15)$$

which is also called *multitaper* spectrogram. The averaging of a few spectrograms forms an effective solution from implementation aspects in contrary to all the steps involving calculation of the Wigner spectrum, transformation to the AS and the corresponding transformation back to the smoothed TF spectrum. The selection of window functions q_k determines the properties of the multitaper spectrogram. A particular choice of multitapers are the Hermite functions which can be computed recursively as

$$\begin{aligned} h_1(t) &= e^{-t^2/2}, \\ h_2(t) &= 2t\, e^{-t^2/2}, \\ h_k(t) &= 2t\, h_{k-1}(t) - 2(k-2) h_{k-2}(t), \quad k = 3 \ \ldots \ K. \end{aligned}$$

As Hermite functions are more localized in the TF plane than Thomson MTs [11, 30], they pose the method of choice in this paper, i.e., we choose $q_k(t) = h_k(t)$, with the corresponding weights $\lambda_k = 1, k = 1 \ldots K$.

The main merit of MT spectrograms is their reduced variance as compared to a single-window spectrogram. The variance of the latter is roughly of order $V[S_x(t,f)] \approx S_x(t,f)^2$. Multitapering using K tapers, however, can lead to a substantial variance reduction. The reason for such improvement in terms of robustness is that the spectrograms from different tapers are uncorrelated, provided the signal satisfies certain properties. Therefore, their average reduces the variance by up to a factor of K, i.e., $V[S_x(t,f)] \approx \frac{1}{K} S_x(t,f)^2$.

In a later section, we will compare methods which rest upon the AS as derived from a single window Hanning spectrogram to those which employ the AS calculated from Hermite MT spectrograms. To facilitate a reasonable comparison between methods based on the Hanning window and those based on Hermite windows, their respective time and frequency concentrations should be related. However, whereas the window length of a Hanning window is well-defined, the Hermite functions are of infinite length and therefore a "window length" is not a reasonable quantity for connecting those windows to each other in a suitable way. Hence, we need to define a measure that relates the two window types. In this paper, we therefore define the time concentration of a window as that time interval in which 99 % of the power is located. A corresponding definition of frequency concentration is to use the frequency interval which contains 99 % of the window's spectral power. For the MT methods, the corresponding concentration values from the window $h_K(t)$ are used, as this is the window with lowest time and frequency concentration in the set of K MTs $h_1, \ldots, h_K$. Note that with a larger value of K, i.e., with more tapers, the time and frequency resolution of the corresponding final estimate will decrease.

Key ingredients to our approach are on the one hand the usage of MTs for spectrogram estimation and on the other hand the transformation to the AS for subsequent feature extraction. The main property of the AS (which also holds for the filtered AS), as already mentioned in the introduction, is its invariance to frequency modulation and time shifts. In fact, for a modulated and time-shifted signal $z(t) = x(t-t_0)e^{i2\pi f_0 t}$ one has $|A_z(\nu,\tau)| = |A_x(\nu,\tau)|$, [28]. This property is desirable in many applications related to acoustic signals. As an example, if the comparison of two identical syllables starting at different time points is based on their respective ambiguity spectra, they will indeed be classified as being the same. Analogously, a frequency modulated syllable, which can be thought of as pronouncing the same syllable in different pitches, will not affect identification of these syllables either.

The popular SPCC approach, which cross-correlates the TF spectra of two signals, shares such robustness to shifts of a signal in time or frequency dimension. The additional advantage of the AS, which will turn out to cause a considerable gain in performance in subsequent examples (as compared to SPCC), can be deduced from Fig. 1. The upper two panels (a) and (b) show two syllable examples from the GRW with conspicuous structural similarities in their time domain representation. Nevertheless, they clearly differ regarding their number of large TF components, which is six for syllable A while being five for syllable B. This mismatch is mirrored in the syllables' spectrograms, see Fig. 1c, d. Therefore, as there is no actual time and frequency location where the two spectrogram images coincide to sufficiently large extend, an SPCC-based syllable comparison will not clearly reveal the striking structural similarities between these signals. In fact, the maximum cross-correlation based on the MT spectrograms is 0.815, not strongly suggesting the substantial

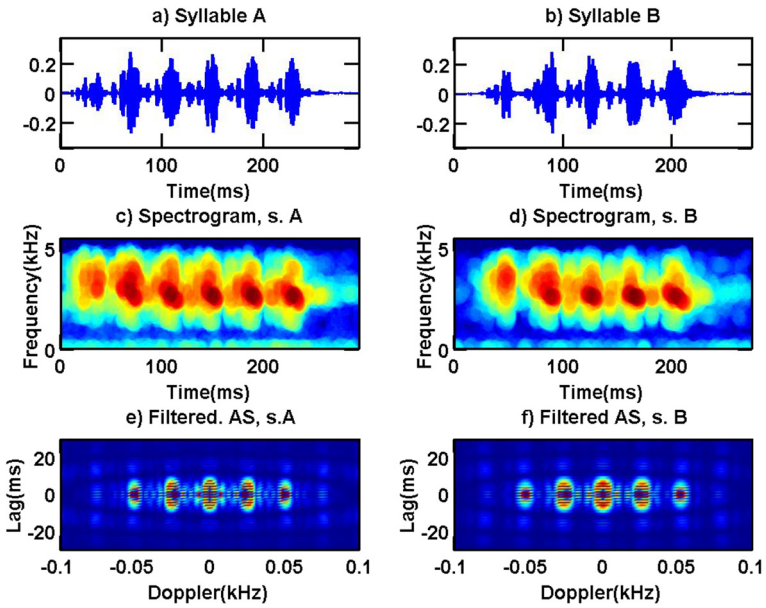

Fig. 1 Example of two syllables: **a** syllable A; **b** syllable B. The corresponding MT spectrograms with window length 6.9 ms, frequency resolution 1770 Hz and $K = 8$;:**c** syllable A; **d** syllable B. The filtered ambiguity spectra: **e** syllable A; **f** syllable B

similarity between these syllables. The syllables' filtered ambiguity spectra, which are depicted in the bottom panels (e) and (f), however, do not reflect these structural discrepancies, but instead closely resemble each other. As opposed to time or spectrogram visualization, the syllables' representation in the ambiguity domain only mirrors distances between the large TF components. These distances are approximately 40 ms (see, e.g., Fig. 1a, in which the components repeat at 70, 110, 150, 190, and 230 ms), corresponding to $1/0.040 = 25$ Hz (and 50 Hz) in Doppler frequency, see panels (e) and (f). Thus, comparison in the ambiguity domain will not be affected by the different numbers of strong components.

3 Feature extraction—singular value decomposition

The singular value decomposition is a low-rank matrix approximation and a popular noise-reduction technique for a data matrix. The decomposition of a matrix $\mathbf{A}$ results in the representation $\mathbf{A} = \sum_{j=1}^{r} \sigma_j \mathbf{u}_j \mathbf{v}_j^H$, where $\mathbf{u}_j, \mathbf{v}_j$ are the singular vectors of unit length and $\sigma_1 \geq \ldots \geq \sigma_r \geq 0$ the singular values. The unit-length vector $\mathbf{v}_1$, i.e., the first right singular vector, maximizes the Euclidean norm $\|\mathbf{Av}\|_2$ and can hence be seen as the vector with unit length which undergoes the maximum amplification under $\mathbf{A}$. Thus, $\mathbf{v}_1$ serves as a crude approximation of the directions of the columns of $\mathbf{A}$. Similarly, $\mathbf{u}_1$ maximizes $\|\mathbf{A}^T\mathbf{u}\|_2$ and serves as an approximation of the row-directions. Hence, if the matrix $\mathbf{A}$ is of low-rank, the vectors $\mathbf{u}_1$, $\mathbf{v}_1$ comprise the major information in $\mathbf{A}$: $\mathbf{u}_1$ captures the frequency-related information while $\mathbf{v}_1$ captures the time-related and the matrix $\tilde{\mathbf{A}}_1 = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^H$, with σ_1 satisfying $\sigma_1 = \|\mathbf{A}^T\mathbf{u}_1\|_2 = \|\mathbf{Av}_2\|_2$, gives a good rank-1 approximation of $\mathbf{A}$.

The ambiguity matrix derived from a song syllable is typically of low rank and therefore predestined for approximation by a small collection of singular vectors. Thus, little information is lost when replacing the AS-based similarity assessment of two syllables with a comparison of the corresponding first left and right singular vectors. More specifically, if $\hat{\mathbf{A}}^{(A)}$ denotes the estimated AS of syllable s_A and $\hat{\mathbf{u}}_1^{(A)}, \hat{\mathbf{v}}_1^{(A)}$ the first pair of singular vectors (with corresponding notation for syllable s_B), similarity between s_A and s_B can be captured by confining the investigation to comparison of $\hat{\mathbf{u}}_1^{(A)}, \hat{\mathbf{v}}_1^{(A)}$ and $\hat{\mathbf{u}}_1^{(B)}, \hat{\mathbf{v}}_1^{(B)}$. Closeness of $\hat{\mathbf{u}}_1^{(A)}, \hat{\mathbf{u}}_1^{(B)}$ then suggests similarity of syllable structures in the frequency domain while closeness of $\hat{\mathbf{v}}_1^{(A)}, \hat{\mathbf{v}}_1^{(B)}$ corresponds to structural resemblance in the time dimension.

As an illustration, the first singular vector pairs of the (estimated) filtered AS of syllables A and B, which are displayed in Fig. 1a, b, are depicted in Fig. 2. The left panel in Fig. 2 illustrates the alikeness of the vectors $\mathbf{u}_1^{(A)}$ and $\mathbf{u}_1^{(B)}$ and the closeness of the pair $\mathbf{v}_1^{(A)}$ and $\mathbf{v}_1^{(B)}$ can be seen in the right panel.

A study of the singular values of the ambiguity spectrum shows that the first pair of singular vectors captures 80 % of the energy in most of the signals. Increasing the number of singular vector pairs to, e.g., 10, would explain

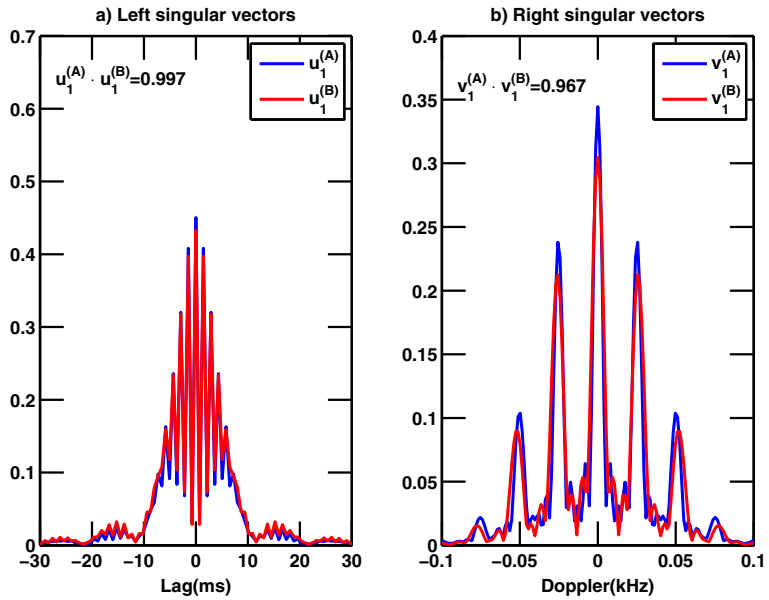

Fig. 2 The corresponding first left and right singular vectors of the filtered ambiguity spectra of syllables A and B in Fig. 1: **a** the left singular vectors, $\mathbf{u}_1^{(A)}$ and $\mathbf{u}_1^{(B)}$; **b** the right singular vectors, $\mathbf{v}_1^{(A)}$ and $\mathbf{v}_1^{(B)}$

approximately 90 % of the variations in the signal. Such gain in captured energy comes, however, at the expense of increased noise and unwanted jitter effects from small differences in similar signals. Restricting signal representation to the first pair is therefore a reasonable choice if the main structure of the signal should be captured. Our investigations including more pairs of singular vectors did not increase the performance.

4 Similarity measures

Denoting the inner product in Euclidean space by $\langle \cdot, \cdot \rangle$, we introduce the following similarity measures

$$\beta_u(s_A, s_B) = |\langle \hat{\mathbf{u}}_1^{(A)}, \hat{\mathbf{u}}_1^{(B)} \rangle|, \tag{16}$$

$$\beta_v(s_A, s_B) = |\langle \hat{\mathbf{v}}_1^{(A)}, \hat{\mathbf{v}}_1^{(B)} \rangle|, \tag{17}$$

where s_A and s_B denote two syllables and $\hat{\mathbf{u}}_1^{(A)}, \hat{\mathbf{v}}_1^{(A)}$ and $\hat{\mathbf{u}}_1^{(B)}, \hat{\mathbf{v}}_1^{(B)}$ their respective pair of first singular vectors. The function β_u thus quantifies similarity of the frequency structures of s_A and s_B whereas time-scale structures are compared in terms of β_v. To assess similarity in time and frequency structure simultaneously, these two measures (which in the following will be referred to as *raw* measures, as opposed to the combined measures introduced below) shall be assembled. To this end we consider and investigate the following three combinations:

$$\beta_{mean}(s_A, s_B) = (\beta_u(s_A, s_B) + \beta_v(s_A, s_B))/2, \tag{18}$$

$$\beta_{min}(s_A, s_B) = \min(\beta_u(s_A, s_B), \beta_v(s_A, s_B)), \tag{19}$$

$$\beta_{max}(s_A, s_B) = \max(\beta_u(s_A, s_B), \beta_v(s_A, s_B)). \tag{20}$$

The normality of singular vectors implies $\beta(s_A, s_A) = 1$. Note, however, that $\beta(s_A, s_B) = 1$ does not suggest equality of syllable s_A and syllable s_B but rather a strong alikeness. In clustering applications, the key issue is to decide whether two syllables are realizations of the same syllable type (and thus should be allocated to the same cluster) or if they arouse from distinct syllable types (and hence should be grouped in different clusters). Due to background noise in recordings and within-individual variability, $\beta(s_A, s_B)$ will rarely be equal to 1 even if s_A, s_B represent the same type of syllable. Thus, the decision on assigning two syllables to the same or to distinct groups will be made based on whether or not $\beta(s_A, s_B)$ exceeds a certain threshold ρ.

5 Syllable detection

The syllable detection approach is divided into two steps. A set of filters is applied in the first step to filter out background noise while the syllables are defined and extracted in the second step based on time distances between amplitude peaks.

Syllables in the initial and final parts of a strophe often have weaker amplitudes than those in the body of the strophe. However, even sections in the middle of a strophe sometimes contain parts with syllables of considerably lower amplitude. Therefore, we have chosen a time-varying adaptive threshold for detection of syllables. This threshold is created by means of two power-smoothing filters (moving averages) of the form

$$P_{filter}(t) = \frac{1}{L} \sum_{t_1=-L/2}^{L/2} x^2(t + t_1), \tag{21}$$

where $x(t)$ represents the time samples of the song and $L+1$ is the length of the filter. The two filters are chosen as one longer P_{long} (default 360 ms), determining a time-varying threshold, and a shorter filter P_{short} (default 90 ms), detecting the actual sample points that belong to a particular syllable. The decision on whether a sample is belonging to a syllable is based on when the level of P_{short} is sufficiently above the level of P_{long},

$$P_{short}(t) > P_{long}(t) + (1 - \frac{l_{sens}}{100}) \cdot \max_t P_{long}(t), \qquad (22)$$

where l_{sens} is the sensitivity of the detector in percentage. After a particular signal section has been declared as a syllable, its start and end time points are extended backwards and forwards (by a default value of ±60 ms) to include the syllable's weaker start and end.

Using a sensitivity of $l_{sens} = 100$ %, Fig. 3a), all minor changes in amplitude level, even those which solely are due to background noise, will be classified as small syllables (see in particular the final part of the signal in Fig. 3a, between 7 and 8 s). Choosing $l_{sens} = 99$ %, Fig. 3b), a slightly less sensitive detector is created that disregards the small amplitude jitters between 7 and 8 s, which are solely due to noise. The weaker notes (usually labeled as initial syllables, [31]), at the beginning of the strophe, are encoded as four syllables. However, these are not seen as the syllables of major interest for analysis [31]. Moreover, disturbances of other birds in the strophe, such as the weak sound at 2.7 s, are as well declared as syllables. Thus, a level of 99 % seems still too sensitive, unless the signal is sufficiently clean, i.e., nearly without any wind noise or disturbances from other birds. Using $l_{sens} = 95$ %, Fig. 3c, the four initial syllables will be not be declared as syllables, neither will be the lower-level disturbance from another bird at 2.7 s. Therefore, in this paper, the syllable detection is based on a sensitivity level of $l_{sens} = 95$ %, which gives a sufficiently sensitive detector which at the same time avoids to falsely declare smaller disturbances (arising from background noise) and initial notes as syllables.

The outputs of the two filters, together with the detected samples using $l_{sens} = 95$ %, for a section of the strophe depicted in Fig. 3c, are shown in Fig. 4a. In this example, these default settings result in a good performance of the syllable detection algorithm. The longer filter length (360 ms) is chosen to give a slowly time-varying threshold and to smooth out the variation of the specific syllable power. The total syllable length is assumed to be 50–300 ms with a maximum allowed length of 400 ms, so the 360 ms smoothing will certainly give a reasonable time-varying threshold of any syllable. The shorter filter is by default of length 90 ms, which is reasonably long to smooth out short-term variations, or even empty

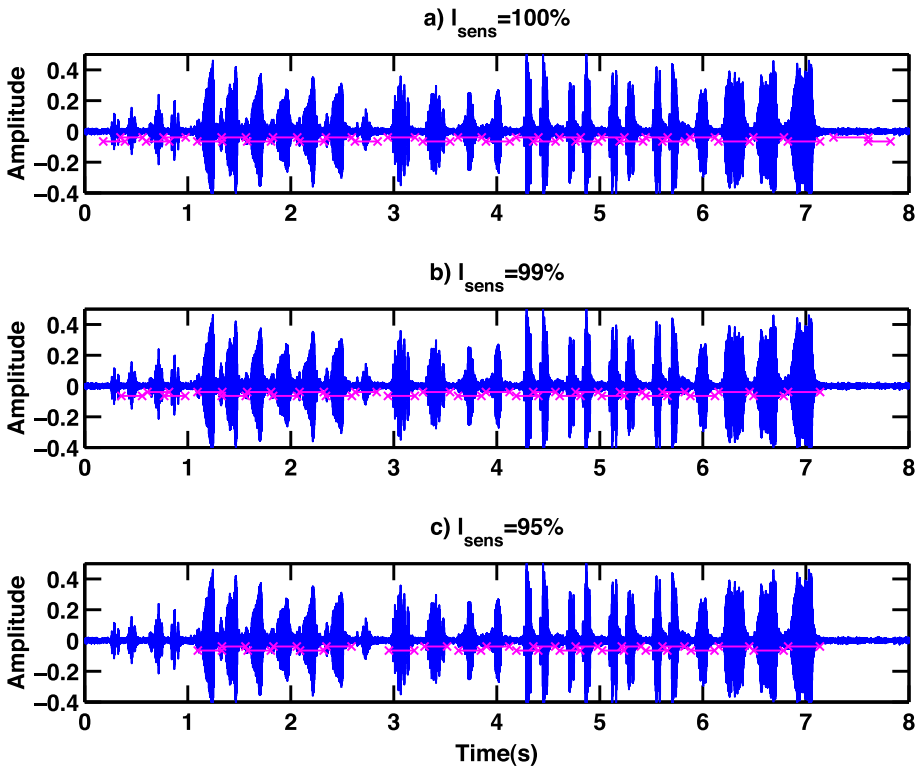

Fig. 3 An example of a strophe with detected syllables (SNR = 20 dB) for **a** $l_{sens} = 100$ %; **b** $l_{sens} = 99$ %; **c** $l_{sens} = 95$ %

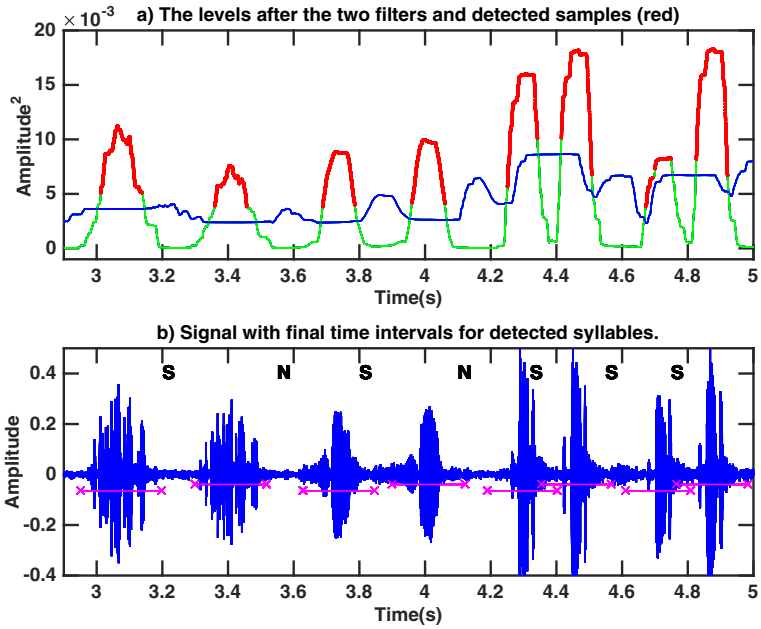

Fig. 4 A part of the strophe in Fig. 3c; **a** The levels of the two smoothing filters, (*blue-long filter, green-short filter*) and detected samples using $l_{sens} = 95$ % (*red*). **b** The corresponding cut syllables including the 60 ms extension backward and forward, and with subsequent pairs of syllables manually labeled as similar (*S*) or non-similar (*N*)

intervals such as, e.g., the silent intervals between two parts of a so called—a syllable composed of two repeated smaller parts. The frequent occurrence of such double syllables is one of the challenges in dealing with songs of the GRW, as it is often unclear whether two detected, closely spaced syllables should be treated as two single or as a double syllable. In the presented detection algorithm, this decision is based on the time distance between the two detected syllables (or syllable parts). If the time distance exceeds a minimum allowed distance (default 60 ms), the detected sounds are declared as two separate syllables. Otherwise, they are assumed to combine into one double syllable.

6 Data presentation and baseline classification

The data under consideration is a 7-min bird-song signal recorded from the Great Reed Warbler under natural outdoor conditions. The bird song has been recorded analogously with a Telinga parabola and microphone and a SONY cassette tape recorder (SONY TC-D5M). The recording is of average quality (with respect to noise and disturbance) and the signal is digitized to a sample frequency of 44.1 kHz, which is subsequently decimated by a factor 4 for the further analysis.

Before the main analysis, the output from the automatic syllable detection step (as described in the previous section) is manually checked for detection errors. In four of the strophes, initial notes were falsely declared as syllables and were therefore removed manually (in total 4×3 syllables). In one strophe a burst of noise was erroneously detected as a syllable and removed as well.

The resulting data to be used for our analysis consists of 362 detected syllables in 28 strophes. A typical strophe section in a GRW song consists of 2–8 repeats of the same syllable type followed by a change to realizations of another syllable structure, see Fig. 3. This characteristic pattern makes it fairly easy to assess whether two subsequent syllables belong to the same or to different types, since the change to another type of syllable is generally quite pronounced. This facilitates a rather straightforward visual (based on the spectrogram and/or the time domain representation of the syllables) and auditive classification of subsequent syllables as being similar (S), i.e., realizations of the same syllable type, or non-similar (N). As an example, the two first syllables in Fig. 4b are labeled as similar, while the pair given by the second and third syllables is marked as non-similar, followed by the as similar declared pair of the third and fourth syllable. This pairwise classification was conducted for all 28 strophes and the resulting labeled data contains 217 subsequent syllable pairs classified as similar and 117 classified as non-similar. This labeling is used as the baseline "truth" for the evaluation of different methods and parameter choices.

The variance of the noise is estimated from the last 1 s of data from each of the 28 strophes, with no signal present. An average of these 28 estimated variances is used as the baseline noise variance $\sigma_N^2 = 0.0002$ of measured data

throughout the paper. Similarly, the mean power of each syllable is computed (without the syllable extension of +− 60 ms), and the average of all the mean powers in a strophe is calculated as the total averaged syllable power, P_{av}. Finally, the signal-to-noise ratio (SNR) of these two values is calculated as

$$SNR = 10\log_{10}\frac{P_{av}}{\sigma_N^2}. \quad (23)$$

The computed SNR of the strophe in Fig. 3 is 20 dB.

Note that due to the special pattern of the GRW song (repeats of a particular syllable type are followed by repetitions of another syllable structure), comparing subsequent syllables depicts a simpler problem than the general approach where all syllables are compared with each other. Defining a baseline truth for method evaluation in the general problem is a much more involved task as it is often difficult to decide (based on listening and visual inspection of spectrograms) whether two syllables, which have been chosen from a song on random basis, are similar and different experts might likely come to different conclusions.

7 Evaluation

With a feature set based on the first singular vectors of the estimated filtered ambiguity spectrum and our proposed similarity measures at hand, we proceed to evaluate our methodology. The target quantities for evaluation are (1) the *similarity rate* $p_S(\alpha)$, i.e., the proportion of correctly classified pairs of similar syllables while accepting $\alpha \cdot 100$ % false positives ($\alpha \cdot 100$ % "non-similar" pairs are misclassified as "similar") and (2) the *non-similarity rate* $p_N(\alpha)$, i.e., the proportion of correctly classified non-similar syllables while allowing for $\alpha \cdot 100$ % false negatives (falsely as "non-similar" classified pairs of "similar" syllables). Here, α is fixed to the value of 0.05.

In the first part, we evaluate the performance of the raw similarity measures β_u and β_v individually, based on different settings for the MT windows. In the second part, we assess the performance of the suggested combined measures for a selection of MT and single window settings.

7.1 Evaluation of raw measures under different window settings

In this section, the performances of the two raw similarity measures are studied. To illustrate the difference between spectrogram-based and ambiguity-based feature representation, we describe syllables via the first singular vector pair of their AS as well as by the first left and right singular vectors from the spectrogram and study the performance of β_u, β_v for both representation settings. Moreover, we consider four approaches for the spectrogram estimation (recall that the AS is just the two-dimensional Fourier transform of a spectrogram), (1) a single Hanning window spectrogram (denoted by H1), and (2) a K-window Hermite MT spectrogram, where K is 2, 4, and 8 (MT2, MT4 and MT8, respectively). Moreover, we combine each spectral estimation technique with a collection of different time and frequency concentrations. In the following, the MT method is labeled as MTK($\Delta t, \Delta f$), where K is the number of tapers, Δt, Δf are the time- and frequency bandwidths, respectively, and the standard single Hanning window is labeled as H1($\Delta t, \Delta f$). Table 1 displays the different constellations of estimation method and TF bandwidths which will be examined.

In the first part, we focus on the performance of β_u and β_v, Eqs. (16 and 17), when applied to feature vectors based on the syllable-specific ambiguity spectra. The results for all considered methods are shown in Fig. 5. In each of the four panels, the x-axis relates to different values of Δt, see Table 1. The upper panels show the similarity rates of the measures β_u and β_v, respectively, while their respective non-similarity rates are illustrated in the two bottom panels. One can observe that - uniformly over all considered methods—the performance of β_u for both similarity and non-similarity rate is satisfactory and moreover robust to changes in the TF bandwidths. On the contrary, its counterpart β_v is highly sensitive to modifications of the TF bandwidths and the performance of all considered methods decreases markedly—both for similarity and non-similarity rates—when the windows are less concentrated. Moreover, it can be seen that the performance of both measures improves as the value of K increases (more tapers are used).

In further investigations, the analysis will be restricted to one Hermite MT and a single Hanning window spectrogram. It is, however, not too obvious which constellations are most qualified. Clearly, best results are given by $K = 8$ MTs, but it is ambiguous which window length should be used for them. In most cases, the window length

Table 1 All different windows and their time and frequency bandwidths

Window	Δt (ms)	Δf (kHz)	Window	Δt (ms)	Δf (kHz)
MT8(53, 0.21)	53.3	0.215	MT2(15, 0.24)	14.9	0.237
MT8(27, 0.43)	26.8	0.431	MT2(7.6, 0.50)	7.62	0.495
MT8(13, 0.88)	13.4	0.883	MT2(3.8, 0.97)	3.81	0.969
MT8(6.9, 1.8)	6.89	1.77	MT2(2.0, 2.0)	1.99	1.96
MT4(39, 0.15)	39.4	0.151	H1(32, 0.065)	31.56	0.065
MT4(20, 0.32)	19.8	0.323	H1(16, 0.11)	15.96	0.108
MT4(10, 0.65)	9.98	0.646	H1(8.0, 0.24)	7.98	0.237
MT4(5.1, 1.3)	5.08	1.31	H1(4.2, 0.47)	4.17	0.474
MT4(2.5, 2.6)	2.54	2.61	H1(2.2, 0.95)	2.18	0.947
MT2(30, 0.13)	29.7	0.129	H1(1.1, 1.8)	1.09	1.83

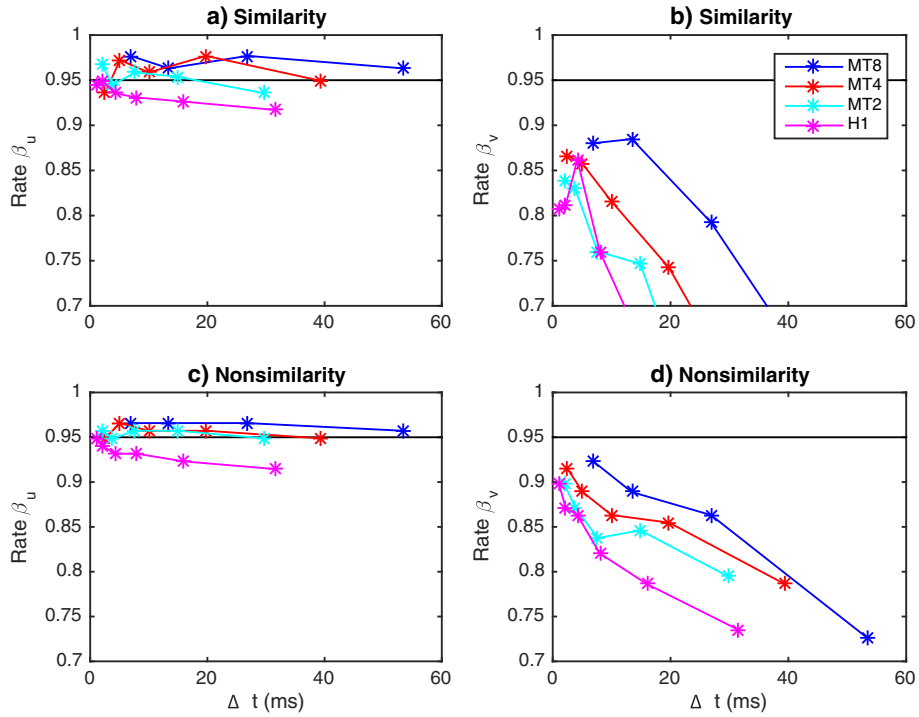

Fig. 5 Similarity rates; **a** β_u; **b** β_v, and non-similarity rates; **c** β_u; **d** β_v, accepting 5 % false positives, based on the AS matrix for all different methods and window lengths defined in Table 1

corresponding to method MT8(13, 0.88) appears to give the best results for MTs, while for the Hanning window spectrogram the choice H1(2.2, 0.95) is considered as most suitable for further analysis. These constellations will in future considerations be referred to as MT8_{AU} when similarity is assessed by filtered ambiguity spectrum and β_u and as MT8_{AV} when instead the measure β_v is employed. Similarly, investigation based on the chosen Hanning window ambiguity spectrum will be referred to as H1_{AU} and H1_{AV}, depending on whether similarity is assessed by β_u or β_v.

In this second part, we repeat the previous investigation, but this time with syllable features extracted from the spectrogram (as opposed to the AS). For this case, the performances of the measures β_u and β_v are presented in Fig. 6. Note that the scales of the y-axes of Fig. 6b, d, which show the results of β_v, is now the whole scale from zero to one. The performance of β_u, Fig. 6a, c, is not as good as for the case when β_u is computed from the AS. The sensitivity to the window length is also higher and the best results are obtained for more concentrated ("shorter") windows with $\Delta t < 10$ ms. Studying β_v in Fig. 6b, d, we see that the classification is rather poor, especially for the rate of non-similarity. This is not an acceptable performance, and therefore, we discard the measure β_v based on the spectrogram for further analysis. As representative for the spectrogram-based SVD, the measure β_u is chosen, both in combination with MT8(13, 0.88) (in the sequel denoted by MT8_{SU}) and in combination with $H1(2.2, 0.95)$ (referred to as H1_{SU}).

To summarize, further investigations will be based on MT8_{AU}, MT8_{AV}, H1_{AU}, H1_{AV}, MT8_{SU}, and H1_{SU}.

7.2 Evaluation of combined measures

In this part we also include the combined measures from Eqs. (18, 19 and 20) in our analysis and investigate which of the similarity measures β_{mean}, β_{max}, β_{min}, β_u, β_v performs best when features are extracted from AS and the computation of AS is based on MT8(13, 0.88). The results are presented in terms of ROC curves, which illustrate the similarity rate $p_S(\alpha)$ as a function of α and are shown in Fig. 7. Here, MT8_{Amean} refers to application of β_{mean} to feature vectors extracted from AS based on MT8(13, 0.88). The remaining notations MT8_{Amin}, MT8_{Amax}, MT8_{AU} and MT8_{AV} have analogous interpretation. Note that only the upper left part of the total ROC figure (total axes 0–1 for both x and y) is shown. It can be seen that the combined measures β_{mean} and β_{min} obtain a similarity rate of 100 % at $\alpha = 0.05$. The performance β_{max} is comparable to that of β_u with similarity rate $p_S(0.05) \approx 0.98$. The application of the raw measure β_v gives a performance well below the other measures.

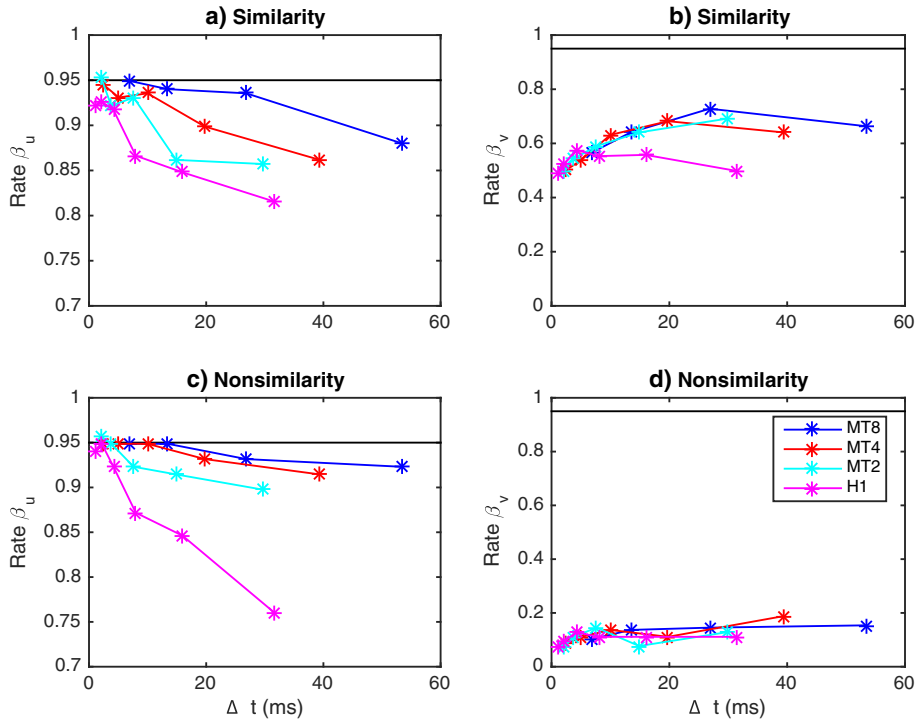

Fig. 6 Similarity rates; **a** β_u; **b** β_v, and non-similarity rates; **c** β_u; **d** β_v, accepting 5 % false positives, based on the spectrogram matrix for all different methods and window lengths defined in Table 1

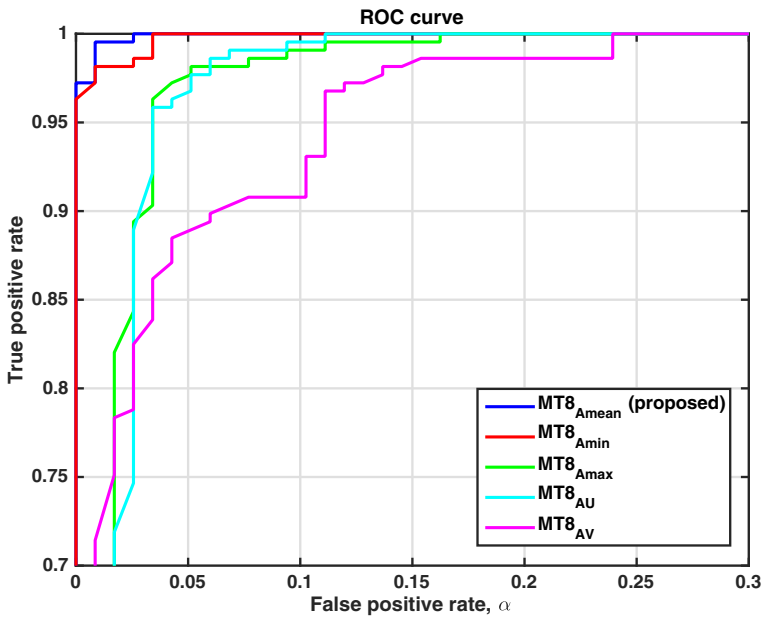

Fig. 7 The ROC curves. The true positive rate (correct classified equal pairs) plotted against the false positive rate (not equal pairs classified as equal) for different similarity measures based on MT8(13, 0.88)

7.3 Comparison with established approaches

Here, we compare more thoroughly the performance of the best method based on the multitaper AS, $\mathrm{MT8}_{Amean}$, to the corresponding Hanning-window approach $\mathrm{H1}_{Amean}$, to the spectrogram-based methods which have been selection in section 7.1 (i.e., to , $\mathrm{MT8}_{SU}$ and $\mathrm{H1}_{SU}$) and moreover to established approaches based on MFCCs and the SPCC. For the MFCC method, the often used implementation by Malcolm Slaney[1] is chosen with eight cepstral coefficients, a 25 ms Hamming window and 90 % overlap between frames. For the SPCC method, we use the single window Hanning spectrogram with time and frequency resolutions 2.18 ms and 947 Hz as defined above.

The results are presented as ROC curves in Fig. 8. When used in combination with β_{mean}, the single Hanning window AS derived from $\mathrm{H1}(2.2, 0.95)$ (i.e., the approach encoded by $\mathrm{H1}_{Amean}$) achieves an almost as reliable result as the method $\mathrm{MT8}_{Amean}$. For both approaches, the similarity rate $p_S(0.05)$ is 100 %. If the raw measure β_u is applied instead of β_{mean} and features are derived from the spectrogram representation, the performance of the single Hanning window approach ($\mathrm{H1}_{SU}$) is comparable to the results if MTs were used for spectrogram estimation ($\mathrm{MT8}_{SU}$) with a still high similarity rate of $p_S(0.05) = 0.95$. A marked drop in performance can, however, be noted for the MFCC- and SPCC-based approaches. These methods give rather unreliable results with similarity rates of 90 % and just above 85 %, respectively.

8 Classification of four predefined syllable classes

The classification of subsequent syllable pairs into "similar" or "non-similar" constitutes the first performance assessment of our methodology. This classification task is, however, beneficial for all algorithms as two subsequent similar syllables are often very similar and therefore alikeness is not too difficult to detect. In the same way, two non-similar syllables are generally sufficiently different such that switches from one syllable-type to another are easily detected by auditive or visual inspection.

Therefore, the methods which have been considered in the previous section (i.e., $\mathrm{MT8}_{Amean}$, $\mathrm{H1}_{Amean}$, $\mathrm{MT8}_{SU}$, $\mathrm{H1}_{SU}$, MFCC and SPCC) are also evaluated for a number of carefully chosen syllables, which are not occurring subsequently. To this end, a subset of 51 syllables has been manually selected. The syllables are chosen in such a way that they can easily be grouped into four different classes, based on visual inspection of their time-domain representation (see Fig. 9) or their spectrograms. For each class, a syllable has been chosen and is displayed in Fig. 10 along with its spectrogram (where the syllable denoted by example 1 was taken from class 1, example 2 from class 2 and so forth). Class 1 and 2 can easily be distinguished based on their time-representation, see the upper two panels in Fig. 9. In contrast, the distinction between syllables belonging to class 3 and 4, if only based on Fig. 9, may be somewhat more involved. However, the spectrogram plots (see Fig. 10) reveal that the main frequency content for syllables of class 1, 2, and 4 is concentrated around 3

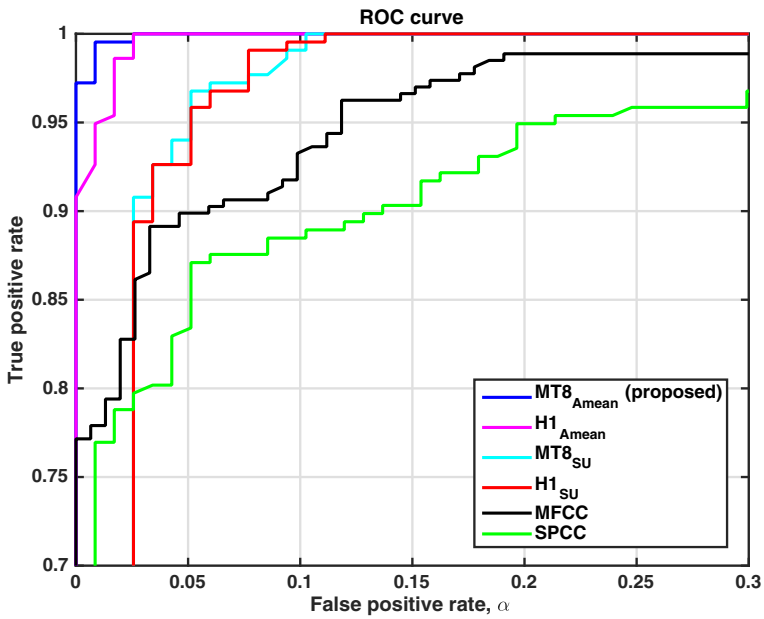

Fig. 8 The ROC curves. The true positive rate (correct classified equal pairs) plotted against the false positive rate (not equal pairs classified as equal) for different similarity measures and methods

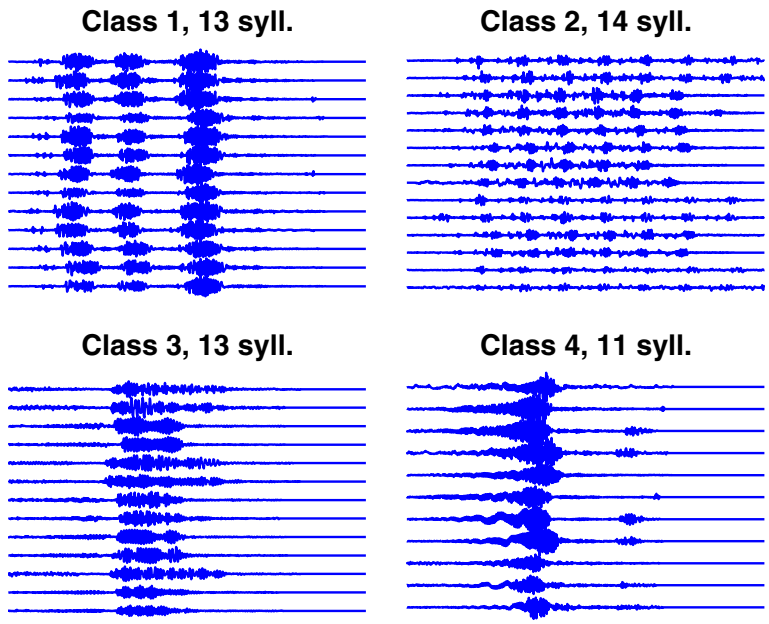

Fig. 9 The syllables of the four chosen classes, SNR = 15 dB

kHz, while the syllables of class 3 have the main concentration around 4 kHz. The difference of the syllables in class 3 is thereby large compared to the syllables of class 4. Therefore, manually separating all 51 syllables based on the respective spectrograms or their time-representations is rather straightforward.

For the considered syllable subset, the average of the mean powers of all 51 syllables resulted in a SNR of 15 dB.

We will investigate and compare the results of the six methods. We first study the correct classification rate for the simpler task of separating syllables from two different classes. This is, for two fixed classes all possible syllable

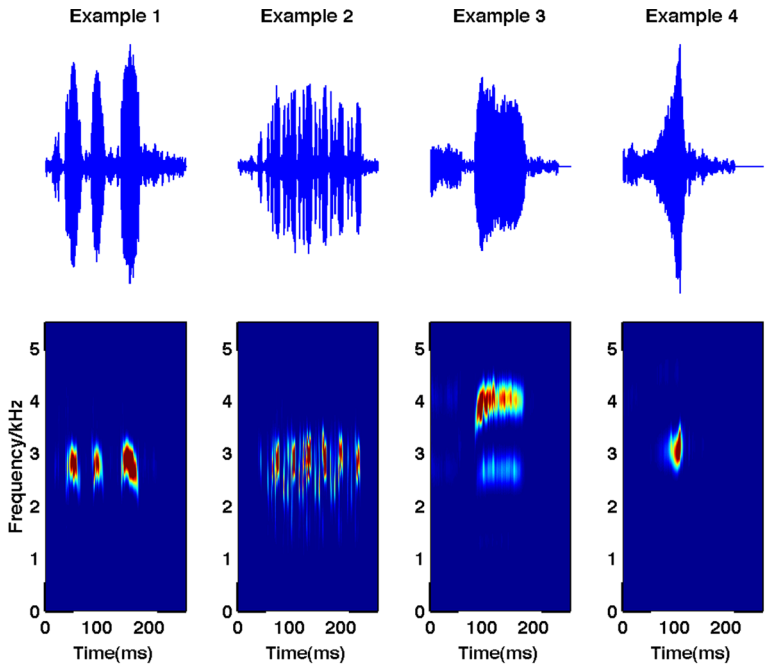

Fig. 10 Example of four different syllables, one from each class in Fig. 9, and their corresponding single window Hanning spectrograms with time resolution 2.18 ms and 947 Hz

pairs are compared with each other. Given two syllable classes C_i, C_j, the total number of comparisons is $N_{total} = N_{between} + N_{within,i} + N_{within,j}$, where $N_{between} = N_i \cdot N_j$ is the number between syllables from different classes and $N_{within,i} = N_i(N_i - 1)/2$ is the number of within-class comparisons for class C_i, similar for class C_j. For class 1 and 2, the number of comparisons within-class will be 169 and between-class will be 182. The results are presented in Table 2, where the similarity rates $p_S(\alpha)$ are shown for all considered methods and $\alpha = 0.05$.

The best results for all class comparisons are given by MT8_{Amean}, closely followed by H1_{Amean}. The results of the spectrogram-based methods MT8_{SU} and H1_{SU} are, however, convincing only for some of the comparisons. For the comparison of class 1 and 2, these methods fail with an achieved similarity rate of only 0.402 and 0.325, respectively. This drop in performance is not surprising as these methods are based on the measure $\mathbf{u}_1$ and therefore entirely disregard the time information contained in the syllables. For the comparison of class 2 and 4 the performances of MT8_{SU} and H1_{SU} are slightly better but still quite unreliable (0.842 and 0.849). The MFCC method performs convincingly for all pairwise class comparisons, apart from the latter (class 3 versus class 4). These two classes include short single syllables, and these do not fulfill the typical structure for which the MFCC method is designed for, i.e., signals, such as speech, with repeating structures suitable for the cepstral decomposition. The results of the SPCC method are promising in all cases, except for the comparisons of class 1 to 2 and class 2 to 4. The insufficient performance for these two class comparisons is connected to the different number of repeats of main components of a syllable in class 2, as already discussed and exemplified in the introduction (see also Fig. 1). There it was noted that cross-correlation of spectrograms will be unreliable if the number of components in the syllables vary.

To further investigate the different methods, we now put our analysis into a more realistic and challenging setting and perform all possible pairwise comparisons between the syllables in the four classes. The number of comparisons within-class is then 302 and between-class is 973. The ROC curves for all methods are computed and depicted in Fig. 11. The true positive rate (y-axis) gives the proportion of correct classification of two syllables belonging to the same class (true positive). The AS-based MT8 method in combination with β_{mean}, i.e., MT8_{Amean}, gives the best result with a rate of 100 % when allowing for 5 % false positives. The similarity rates of H1_{Amean} and SPCC are similar at $\alpha = 0.05$ with $p_S(\alpha) \approx 0.93$. However, both methods are clearly inferior to MT8_{Amean}. The other three methods (MFCC, MT8_{SU} and H1_{SU}) surrender to the complexity of the analysis task, revealing a poor performance of $p_S(0.05) = 0.74$ and $p_S(0.05) \approx 0.05$. In order for those methods (as well as for SPCC) to achieve a similarity rate of $p_S(\alpha) \geq 0.95$, one has to concede to a false positive acceptance rate of as large as $\alpha = 0.2$. Thus, without accepting a very high rate of false positives, it will be difficult to find a reasonable rate of correct classification for these methods.

To investigate the performance for higher noise levels, white Gaussian noise realizations with variance σ_{ext}^2 are added to all syllables of the four classes and the new SNR is defined as

$$SNR = 10 \log_{10} \frac{P_{av}}{\sigma_N^2 + \sigma_{ext}^2}, \tag{24}$$

where P_{av} and σ_N^2 are defined as previously. For different methods, the similarity rates $p_S(0.05)$ of accepting a false rate of 5 % are depicted in Fig. 12a) as a function of the SNR. Here we chose SNRs ranging from 15 dB (SNR of the measured signal) to -2 dB. The methods depicted in that figure are the same as those in Fig. 11. The results show that MT8_{Amean} is reliable with a correct rate of 100 % for SNR values up to 3 dB, where H1_{Amean} achieves only around 95 % of correct classifications and the SPCC reaches 90 %. Between 3 and -2 dB both MT8_{Amean} and H1_{Amean} show a similar result, achieving a similarity rate of about 95 %. The four example syllables in Fig. 10 are shown in Fig. 12b) for SNR = 0 dB.

Table 2 Rates for correct classification of two syllables belonging to the same class accepting 5 % false positives

Class comp.	$MT8_{Amean}$	$H1_{Amean}$	$MT8_{SU}$	$H1_{SU}$	*MFCC*	*SPCC*
1 – 2	1.0	0.994	0.402	0.325	0.905	0.846
1 – 3	1.0	1.0	1.0	1.0	0.987	1.0
1 – 4	1.0	1.0	1.0	1.0	0.932	1.0
2 – 3	1.0	1.0	1.0	1.0	1.0	1.0
2 – 4	1.0	1.0	0.842	0.849	0.931	0.842
3 – 4	1.0	1.0	1.0	1.0	0.421	1.0

9 Conclusions

In this work, a novel feature set for low-dimensional signal representation is suggested that is designed for the analysis of non-stationary signals with complex variation in time and frequency. The features for signal representation are given by the first pair of singular vectors from the MT ambiguity spectrum, which ensures robustness to noise, and shifts in time, frequency and amplitude. For classification or and clustering purposes of a signal (e.g., a bird song), a collection of similarity measures are proposed. These are compared and evaluated on the basis of an outdoor recording of a wild male Great Reed Warbler, being a bird species with complex song structure. Moreover, it is shown that the suggested signal representation along with a specific combined similarity measure (which uses

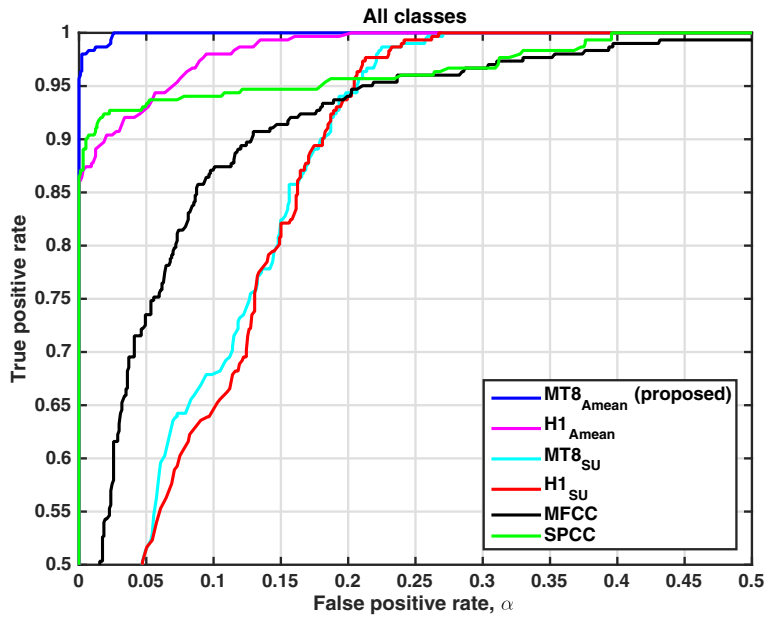

Fig. 11 The ROC curves for correct classification of two syllables belonging to the same class for all methods

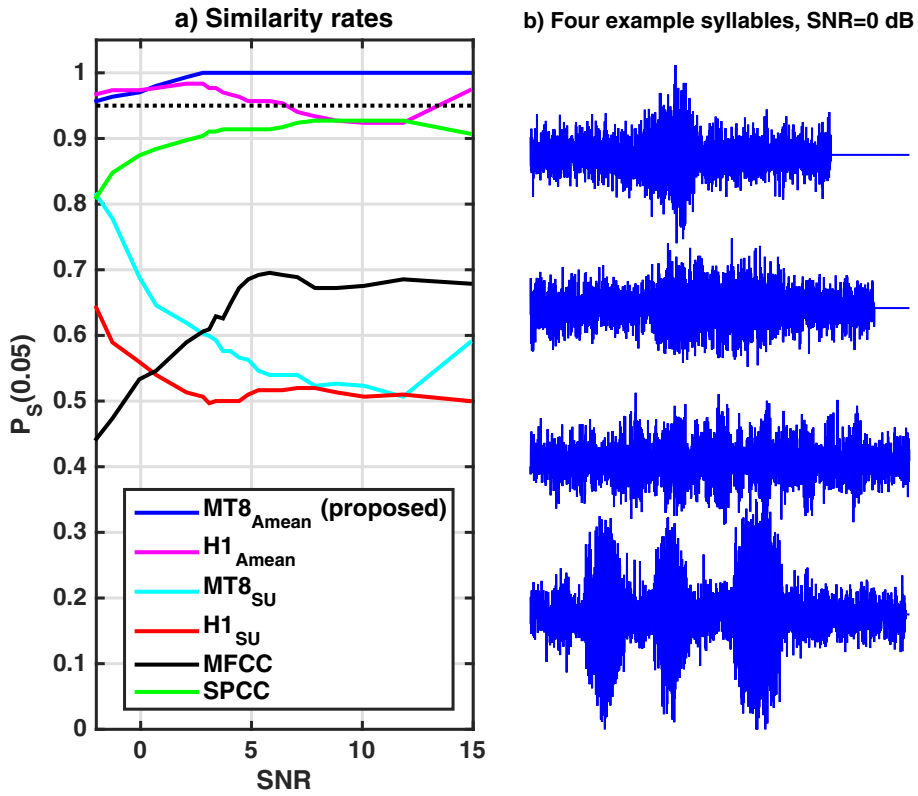

Fig. 12 Simulation with added white Gaussian noise. **a** The similarity rates $p_S(0.05)$ for different SNR. **b** The four syllables in Fig. 10 with added noise giving SNR = 0 dB

an average of the inner product of right singular vectors and of the left singular vectors) clearly outperforms other well-known methods (SPCC and MFCC) in the example analysis of the bird-song data.

Our methodology is also compared to a similar approach where the AS is replaced by the spectrogram for feature extraction, and it could be is observed that switching to the spectrogram comes along with a marked evident decrease in performance. Furthermore, we compared calculation of the spectrograms by means of based on (Hermite) MTs to spectrograms based on a single Hanning window and concluded that MTs increase the performance in a classification task, both for AS-based and for spectrogram-based feature representation.

Endnote

[1] https://engineering.purdue.edu/~malcolm/interval/1998-010/

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
Thanks to the eSSENCE academy for funding and to the Department of Biology, Lund University, for data collection.

Author details
[1]Lund university, Mathematical Statistics, Centre for Mathematical Sciences, Lund, Sweden. [2]University of Copenhagen, Department of Mathematical Sciences, Copenhagen, Denmark.

Received: 14 August 2015 Accepted: 17 May 2016
Published online: 31 May 2016

References

1. ERA Cramer, Measuring consistency: spectrogram cross-correlation versus targeted acoustic parameters. Bioacoustics: Int. J. Anim. Sound Recording. **22**(3), 247–257 (2012)
2. S Keen, JC Ross, ET Griffiths, M Lanzone, A Farnsworth, A comparison of similarity-based approaches in the classification of flight calls of four species of north american wood-warblers (parulidae). Ecol. Informatics. **21**, 25–33 (2014)
3. S Fagerlund, UK Laine, New parametric representations of bird sounds for automatic classification, *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8247–8251 (2014). http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6855209&isnumber=6853544
4. CD Meliza, SC Keen, DR Rubenstein, Pitch- and spectralbased dynamic time warping methods for comparing field recordings of harmonic avian vocalizations. J. Acoust. Soc. Am. **134**(2), 1407–1415 (2013)
5. O Tchernichovski, TJ Lints, S Deregnaucourt, A Cimenser, PP Mitra, Studying the song development process. rationale and methods. Ann. NY Acad. Sci. **1016**, 348–363 (2004)
6. P Somervuo, Härmä, S Fagerlund, Parametric representations of bird sounds for automatic species recognition. IEEE Trans. Audio Speech Lang. Process. **14**(6), 2252–2263 (2006)
7. X Zhang, Y Li, Adaptive energy detection for bird sound detection in complex environments. Neurocomputing. **155**, 108–116 (2015)
8. D Hasselquist, S Bensch, T von Schantz, Correlation between male song repertoire, extra-pair paternity and offspring survival in the great reed warbler. Nature. **381**, 229–232 (1996)
9. E Węgrzyn, 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. **151**, 255–267 (2010). doi: 10.1007/s10336-009-0451-x
10. DJ Thomson, Spectrum estimation and harmonic analysis. Proc. IEEE. **70**(9), 1055–1096 (1982)
11. I Daubechies, Time-frequency localization operators: a geometric phase space approach. IEEE Trans. Information Theory. **34**(4), 605–612 (1988)
12. B Jokanovic, MG Amin, YD Zhang, F Ahmad, Multi-window time-frequency signature reconstruction from undersampled continuous-wave radar measurements for fall detection. IET Radar, Sonar Navigation. **9**(2), 173–183 (2015)
13. M Hansson-Sandsten, Optimal estimation of the time-varying spectrum of a class of locally stationary processes using Hermite functions. EURASIP J. Adv. Signal Process (2011). Article ID 980805
14. Orović, Stanković, M Amin, A new approach for classification of human gait based on time-frequency feature representations. Signal Process. **91**(6), 1448–1456 (2011)
15. P Wahlberg, M Hansson, Kernels and multiple windows for estimation of the Wigner-Ville spectrum of gaussian locally stationary processes. IEEE Trans. Signal Process. **55**(1), 73–87 (2007)
16. M Hansson-Sandsten, M Tarka, J Caissy-Martineau, B Hansson, D Hasselquist, SVD-based classification of bird singing in different time-frequency domains using multitapers, *Signal Processing Conference, 2011 19th European*, 966–970 (2011). http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7073944&isnumber=7069645
17. M Hansson-Sandsten, Classification of bird song syllables using singular vectors of the multitaper spectrogram, *Signal Processing Conference, 2015 23rd European*, 554–558 (2015). http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7362444&isnumber=7362087
18. B Boashash, NA Khan, T Ben-Jabeur, Time-frequency features for pattern recognition using high-resolution TFDs: A tutorial review. Digital Signal Process. **40**, 1–30 (2015). http://dx.doi.org/10.1016/j.dsp.2014.12.015
19. RJ Barry, FM de Blasio, EM Bernat, GZ Steiner, Event-related EEG time-frequency PCA and the orienting reflex to auditory stimuli. Psychophysiology. **52**, 555–561 (2015)
20. Z Yu, Y Sun, W Jin, A novel generalized demodulation approach for multi-component signals. Signal Process. **118**, 188–202 (2016)
21. DD Lee, HS Seung, Learning the parts of objects by non-negative matrix factorization. Nature. **401**(6755), 788–791 (1999)
22. R Hennequin, R Badeau, B David, NMF with time-frequency activations to model non-stationary audio events. IEEE Trans. Audio Speech Lang. Process. **19**(4), 744–753 (2011)
23. B Ghoraani, S Krishnan, Time-frequency matrix feature extraction and classification of environmental audio signals. IEEE Trans. Audio Speech Lang. Process. **19**(7), 1071–1083 (2011)
24. B Ghoraani, Selected topics on time-frequency matrix decomposition analysis. J. Pattern Recognit. Intell. Syst. **1**(3), 64–78 (2013)
25. H Qiao, New SVD based initialization strategy for non-negative matrix factorization. Pattern Recognit. Lett. **63**, 71–77 (2015)
26. D Groutage, D Bennink, Feature sets for nonstationary signals derive from moments of the singular value decomposition of cohen-posch (positive time-frequency) distributions. IEEE Trans. Signal Process. **48**(5), 1498–1503 (2000)
27. M Große Ruse, D Hasselquist, B Hansson, M Tarka, M Sandsten, Automated analysis of song structure in complex birdsongs. Animal Behav. **112**, 39–51 (2015). http://dx.doi.org/10.1016/j.anbehav.2015.11.013
28. B Boashash, *Time Frequency Signal Analysis and Processing; A Comprehensive Reference*, 1st edn. (Elsevier Ltd, The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK, 2003)
29. L Cohen, *Time-Frequency Analysis*, 1st edn. (Prentice-Hall Inc., Upper Saddle River, NJ, USA, 1995)
30. M Bayram, RG Baraniuk, Multiple window time-frequency analysis, *Time-Frequency and Time-Scale Analysis, 1996., Proceedings of the IEEE-SP International Symposium on*, 173–176 (1996). http://ieeexplore.ieee.org.ludwig.lub.lu.se/stamp/stamp.jsp?tp=&arnumber=547209&isnumber=11466
31. K Leniowski, E Węgrzyn, Organization, variation in time, and impacting factors in the song strophe repertoire in the great reed warbler (acrocephalus arundinaceus). Ornis Fennica. **90**, 129–141 (2013)