A SVD-BASED CLASSIFICATION OF BIRD SINGING IN DIFFERENT TIME-FREQUENCY DOMAINS USING MULTITAPERS

Maria Hansson-Sandsten, Maja Tarka, Jessica Caissy-Martineau, Bengt Hansson, Dennis Hasselquist*

*Mathematical Statistics,
Centre for Math. Sciences, Lund University,
Box 118, SE-221 00 Lund, Sweden,
phone: + (46) 46 22 249 53, fax: + (46) 46 22 246 23,
email: sandsten@maths.lth.se.

Molecular Ecology and Evolution Lab,
Department of Biology, Lund University,
Ecology Building, SE-223 62 Lund, Sweden,
phone: + (46) 46 22 249 96
email: bengt.hansson@zooekol.lu.se.

ABSTRACT

In this paper, a novel method for analysing a bird's song is presented. The song of male great reed warblers is used for developing and testing the methods. A robust method for detecting syllables is proposed and a classification of those syllables as compared to reference syllables is done. The extraction of classification features are based on the use of singular vectors in different time-frequency domains, such as the ambiguity and the doppler domains, in addition to the usual sonogram. The analysis is also made using multitaper analysis where the Welch method and the Thomson multitapers are compared to the more recently proposed locally stationary process multitapers.

1. INTRODUCTION

A bird's song is used under several circumstances. It may be used as an identification tool, serving as a recognition signal to indicate the individual, the kinship and the species. The song is also often used in male-male competition and may play a role into the mate choice of a female. Having several functions related to the biology and ecology of birds, their songs have been greatly studied. The studies are however impaired by the lack of methods which would automatically and more objectively analyze the song structure. While some methods have been developed to create a symmetric pairwise similarity comparison within a single song such as Song Analysis Pro, [1], the possibility to compare two songs, whether to evaluate the song properties of the same bird recorded at several occasions or to create a population study, is still lacking. The properties in the song, which cause variations and similarities from e.g., one year to another and songs from the same and different populations, are still an unexplored field, and call for modern tools.

One of the bird species that has been thoroughly studied in terms of song complexity is the great reed warbler, which is the largest warbler species in Europe and a species with exceptional song capacity. A long-term study of a great reed warbler population in Sweden is ongoing, and a main aim is to understand the role of the song in an ecological and evolutionary context. The song of male great reed warblers is used in this project for developing and testing the methods.

Time-frequency analysis of non-stationary processes is an important area with many applications and a significant number of methods have been suggested over the years. For analysis of bird singing, the sonogram, or the time-frequency spectrum has been used, probably because the time-frequency representation is intuitive. However, in time-frequency analysis, four possible representations can be found, the instantaneous auto-correlation function, the ambiguity domain and the doppler domain in addition to the Wigner domain representation (sonogram), [2]. To actually be able to identify similarities and differences in the bird singing, investigation of the properties in other domains than the time-frequency representation could be valuable.

The recording of bird singing, if not made in a laboratory, is difficult. The environment is often noisy, e.g., from different wind conditions, and in addition to this, songs from other birds further away, might disturb the recording. This calls for robust analysis algorithms. Computationally efficient robust algorithms for time-frequency analysis can be found using e.g., multitapers. The phrase multitaper was originally introduced by Thomson, [3], for the case of stationary processes with smooth spectra. The properties that give uncorrelated spectra come from the windows and not from the time-division of data. All data samples for all windows as are totally overlapping and thereby more of the information in data is used than, e.g., the Welch or the WOSA method, [4]. More recently, smoothed Wigner spectra have been shown to approximate multitaper spectrograms, where the multitapers and weights correspond to the eigenvectors and eigenvalues of a time-lag kernel, [5]. Multitaper decomposition of time-lag kernels have been analyzed from several aspects, for existing kernels but new multitaper techniques for non-stationary signal analysis have also been proposed, [6, 7].

A locally stationary process (LSP) has a covariance function which is a multiplication of a covariance function of a stationary process and a time-variable function, [8]. The process is non-stationary with properties suitable for modeling measured signals that e.g., have a transient amplitude behavior. In this paper, the LSP multitapers, [7], is compared to the more well known Thomson multitapers, [3], and Welch method, [4] for analysis of bird singing. A classification comparison is also made, using different time-frequency representations, such as the Wigner domain, the ambiguity domain and the doppler domain. In addition to this, we suggest a new, robust, extraction algorithm for syllable detection in the bird singing. In section 2, a brief description of the different time-frequency domains are given and in section 3, the spectrogram decomposition into multitaper spectra of time-frequency kernels is presented. Section 4 describes the syllable detection and feature extraction in the different domains. Analysis and results from a noisy strophe is presented in section5 and finally the conclusions is given in section 6.

Thanks to the Swedish Research Council for funding.

2. TIME-FREQUENCY ANALYSIS

For a non-stationary process we can define the **instantaneous autocorrelation function**, (IAF),

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

of the zero-mean process $x(t)$ where $E[*]$ is the expected value. We have two variables t and τ and an extension of the Wiener-Khintchine theorem to the **time-varying spectral density** and Fourier transforming from τ to f, gives $W_x(t,f) = \mathscr{F}_\tau r_x(t,\tau)$, also called the **Wigner spectrum**, [2]. In the quadratic class we find the **smoothed Wigner spectrum** as the 2-dimensional convolution of the Wigner spectrum and the **time-frequency kernel**, $\Phi(t,f)$, as

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

The Fourier transform of the IAF in the variable t gives $A_x(\nu,\tau) = \mathscr{F}_t r_x(t,\tau)$, which is called the **Ambiguity spectrum**. The multiplication of an **ambiguity domain kernel**, $\phi(\nu,\tau)$, and the ambiguity spectrum is

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

giving the filtered ambiguity spectrum. We can also reformulate this as a convolution in the variable t in the time-lag domain,

$$r_x^Q(t,\tau) = r_x(t,\tau) * \rho(t,\tau), \tag{4}$$

where $\rho(t,\tau)$ is the **time-lag kernel**.

We may also take the Fourier transform of the ambiguity spectrum in τ or the Fourier transform of the Wigner spectrum in t or, $D_x(\nu,f) = \mathscr{F}_\tau A_x(\nu,\tau) = \mathscr{F}_t W_x(t,f)$, which is called the **Doppler spectrum**. A convolution, in the variable f, of the Doppler spectrum and the **doppler kernel**, $\kappa(\nu,f)$, is expressed as

$$D_x^Q(\nu,f) = D_x(\nu,f) * \kappa(\nu,f). \tag{5}$$

These four different domains for representation of a time-varying signal is shown in a schematic overview is given in Figure 1.

3. SPECTROGRAM DECOMPOSITION OF TIME-FREQUENCY KERNELS

The connection between a multitaper spectrogram and a smoothed Wigner spectrum is found using the following approach, [5]. The multitaper spectrogram is defined as

$$\begin{aligned} S_x(t,f) &= \sum_{k=1}^{K}\alpha_k|\int_{-\infty}^{\infty}h_k^*(t-t_1)x(t_1)e^{-i2\pi f t_1}dt_1|^2, \quad (6)\\ &= \sum_{k=1}^{K}\alpha_k(\int_{-\infty}^{\infty}h_k^*(t-t_1)x(t_1)e^{-i2\pi f t_1}dt_1)\times\\ &\quad \times(\int_{-\infty}^{\infty}h_k(t-t_2)x^*(t_2)e^{i2\pi f t_2}dt_2). \end{aligned}$$

With $t_1 = t' + \frac{\tau}{2}$ and $t_2 = t' - \frac{\tau}{2}$,

$$\begin{aligned} S_x(t,f) &= \sum_{k=1}^{K}\alpha_k\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}x(t'+\frac{\tau}{2})x^*(t'-\frac{\tau}{2})\times\\ &\quad \times h_k(t-t'-\frac{\tau}{2})h_k^*(t-t'+\frac{\tau}{2})e^{-i2\pi f\tau}d\tau dt'. \end{aligned}$$

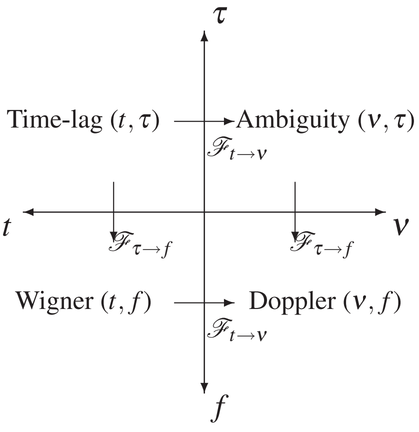

Figure 1: The four possible domains in time-frequency analysis.

We identify the instantaneous autocorrelation function as $r_x(t,\tau) = x(t+\frac{\tau}{2})x^*(t-\frac{\tau}{2})$ and the time-lag kernel

$$\rho(t,\tau) = \sum_{k=1}^{K}\alpha_k h_k(t+\frac{\tau}{2})h_k^*(t-\frac{\tau}{2}), \tag{7}$$

giving the quadratic class of time-frequency distributions as

$$\begin{aligned} S_x(t,f) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}r_x(t',\tau)\rho(t-t',\tau)^*e^{-i2\pi f\tau}dt'd\tau\\ &= W_x(t,f)**\Phi(t,f) = W_x^Q(t,f), \quad (8) \end{aligned}$$

Defining

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

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

$$\rho^{rot}(t_1,t_2) = (\rho^{rot}(t_2,t_1))^*,$$

then solving the integral

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

results in eigenvalues λ_k and eigenfunctions $q_k(t)$ which form a complete set and can be used as weights, α_k, and multitapers, $h_k(t) = q_k(t)$, $k = 1 \ldots K$, in Eq. (6).

4. SYLLABLE ANALYSIS

The analysis of syllables requires a number of steps, automatic detection of a syllables, feature extraction and classification. In this section, brief descriptions of the algorithms are given. The recording of the bird singing in the example is made in the natural environment. The recorded signal is sampled with $f_s = 44.1$ kHz and decimated a factor 4. A strophe (partly seen in Figure 2), $x_s^r(n)$, $n = 0\ldots$, is cut into the syllables, $x_j^r(n)$, $n = 0\ldots N_j-1$, where N_j is varying according to the length defined by the detection algorithm.

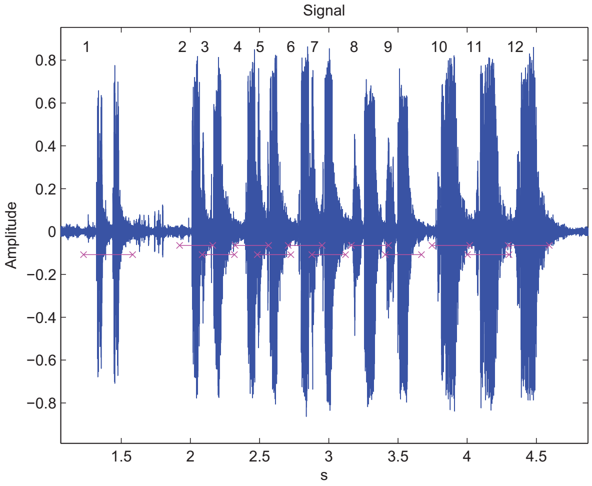

Figure 2: An example of a strophe with detected and numbered syllables.

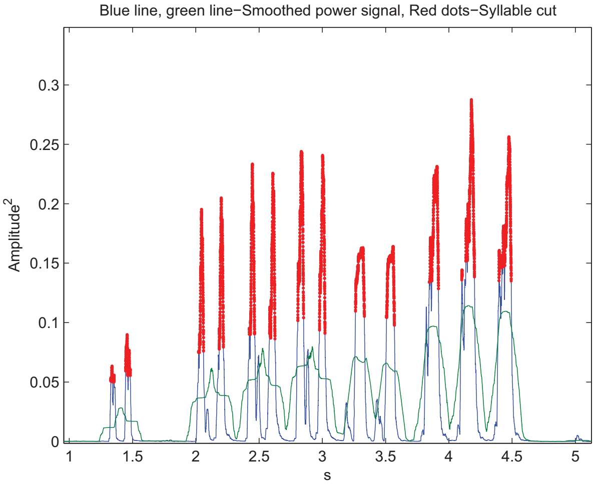

Figure 3: The long filtered power (green line) and the short filtered power (blue line). The red line, which is used for the further detection of syllables show where the blue line is extending the green line with 30 % of the maximum power found in the strophe.

4.1 Automatic detection of syllables

The recorded strophes are very noisy, and methods used for the detection need to be robust. A pre-filtering is made using a bandpass FIR-filter of length 100 with cut-off frequencies 700 and 4500 Hz. An example of a filtered strophe is seen in Figure 2. Using a fixed threshold will cause difficulties as the recording amplitude might vary a lot, even during a strophe, caused by e.g., changes in the wind direction. Therefore, we have chosen an adaptive threshold that is not sensitive to noise, nor to changing amplitude. We suggest a method using two filters, a longer and a shorter smoothing filter, where the power, (square of the signal amplitude) is smoothed. The longer filter is of length 200 ms and the shorter of length 20 ms. In Figure 3, an example of the different powers from the filtering procedure is seen, the green line is the smoothed power from the long filter and the blue line from the short filter. We use the green line as adaptive threshold, catching slow power changes in the signal. The short filter (blue line) catches the fast changes and the decision is based on when this signal is a certain level above the adaptive threshold (green line). This level is chosen as a certain percentage of the maximum instantaneous power found in the signal, and is marked by red dots in Figure 3. As default value we use 30% of the maximum power as the amount for the green line to be recognized as detected power of a possible syllable. The default value of 30% should be changed to a lower value for recordings without too much varying noise. Weaker parts of the strophes might then also be included for analysis. Using this measure, we also avoid including syllables of other birds singing further away, which will show up as a similar but weaker signal, in or between the strophes.

When the samples of a possible syllable is identified, we check if the time distance between the samples exceeds a certain limit (default 115 ms), which is defined as the minimum distance between two different syllables. If the interval width between two sequential detected power samples (red color) is smaller than 115 ms, the corresponding parts of the time signal are included in the same syllable, and all parts of the signal is used for further analysis. This part is tricky as many syllables show up in pairs and it is not clear if these should be treated as two syllables or a so-called *double syllable*. Using a default value of 115 ms will include double syllables in the analysis, treated as one signal in further analysis. The minimum time distance could however, be changed to a smaller value, defining shorter parts of the strophe as syllables. The start and end time points of the defined syllable are extended backwards and forward to include the weaker start and end of the signal, (default ±100 ms), defining the syllable $x^r_j(n)$, $n = 0 \dots N_j - 1$, where N_j varies between $2000 - 4000$ samples using the decimated sample frequency $f_s = 11.025$ kHz.

4.2 Feature extraction and classification

From the defined syllables, $x^r_j(n)$, the feature extraction is made. Time-frequency analysis is used and different multitaper techniques are applied for the computation of the spectrogram, the ambiguity spectrum and the doppler spectrum. For all analysis, the real valued data $x^r_j(n)$ is Hilbert transformed to the analytic time domain signal, $x_j(n)$. The time-frequency, the ambiguity and doppler domain computations are made using a fast Fourier transform (FFT) length of $L = 1024$ and a window-shift for the different computations of 16. The window lengths of the different multitapers applied are 23 ms, (256 samples). An example of a single syllable and the resulting representations in the different domains are seen in Figure 4. We have chosen to compare features from these three domains for the classification of the different syllables in a strophe. However, instead of comparing all information in the two-dimensional matrices for different syllables in the classification, the features are extracted as the first singular vectors from the singular value decomposition (SVD) of the absolute values of the time-frequency domain, the ambiguity domain and the doppler domain respectively. The SVD extracts the basic information of the matrix and using just the first singular vectors reduces the influence of the smaller changes and noise. The basic information included in the singular vectors could be seen as a represen-

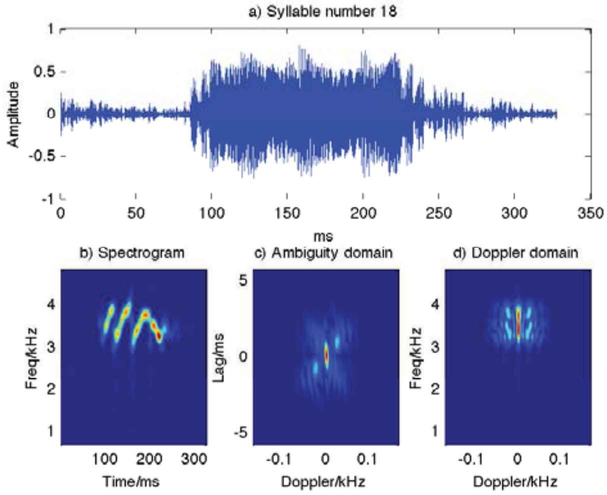

Figure 4: Example of a syllable in a) the time domain, b) the time-frequency domain, c) the ambiguity domain, d) the doppler domain, computed from the LSP multitapers.

tation that also differ between the domains. E.g., studying the time-frequency domain, a syllable and the same syllable modulated in frequency will not give comparable singular vectors for analysis, i.e., a classification based on this will indicate a difference. Similarly, time differences will show up in the other singular vector. In the ambiguity domain, however, neither a difference in modulation of the frequency nor a difference in time will show up, the ambiguity domain representations are equal for time modulations as well as frequency modulations of analytic signals, [2]. In the Doppler domain, a time modulation will give equal representation but not a frequency modulation.

We assume a number of analysis syllables J_a and a number of reference syllables J_r and classify analysis syllable j_a to belong to the reference syllable j_r where the minimum of the sum of the total squared error of analysis vectors and the reference vectors is found, i.e.,

$$e^z_{j_r}(j_a) = \sum_n (U^z_{j_a}(n) - U^z_{j_r}(n))^2 + \sum_n (V^z_{j_a}(n) - V^z_{j_r}(n))^2, \tag{10}$$

where $U^z_{j_a}(n)$ and $V^z_{j_a}(n)$ are the singular vectors from the analysis syllable and $U^z_{j_r}(n)$ and $V^z_{j_r}(n)$ are the singular vectors from the reference syllable. The variable z is equal to W (time-frequency), A (ambiguity) or D (doppler), depending on the domain of the analysis.

5. ANALYSIS AND RESULTS

As a first step a successful algorithm should identify similar syllables that are repeated after each other in a strophe as belonging to the same class. In the example presented here we use the strophe partly presented in Figure 1 and extract the first syllable in a repetition of syllables as the reference. We extract syllable number 1,2,8,10,14, 18 and 21 of in total 23 detected syllables (note that just the 12 first are seen in Figure 1). In the analysis, we are now interested to see if the following syllables on each reference syllable will be classified to the correct class. For a successful classification of repetition of syllables, number 3,4,5,6,7 should belong to

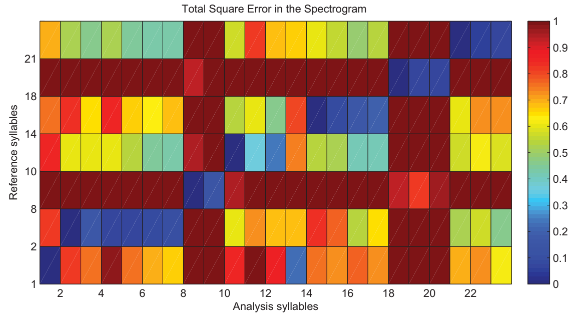

Figure 5: The total square error $e^W_{j_r}(j_a)$ of the time-frequency domain using the LSP multitapers. The blue color shows small error compared to the reference syllable (y-axis) where the red color shows large error.

syllable 2, number 9 to syllable 8, number 11, 12 to 10, number 15,16,17 to 14, number 19, 20 to 18, and finally number 22, 23 to 21. We also assume that syllable 13 might be similar to syllable 1.

For the computation, a single Hanning window, the Welch method, [4], the Thomson multitapers, [3], and the LSP multitapers, [7], are applied. In all cases, the total window lengths are $N_w = 256$ samples, (23 ms). For the Welch method, [4], a number of K=4 Hanning windows are used with the usual overlap of 50 % between windows. A number of K=4 equally weighted spectrograms using Thomson multitapers are also applied where the resolution bandwidth for computation of the tapers are $B = (K+3)/N_w \approx 0.0273$, which gives a resolution of about 0.3 kHz in the analysis. The LSP multitapers, optimal to a LSP of two Gaussian functions, have been shown to approximate a set of Hermite functions fairly well, [7], which makes them more suitable for implementation. The LSP defined by two Gaussian functions, could also be a reasonable model for the syllables as they usually arise and decrease smoothly. In this analysis, the weighting applied to the set of Hermite function spectrograms, are $\alpha_k = 0.371, 0.255, 0.169, 0.108, 0.0638, 0.0333$, for $k = 1 \ldots 6$, which corresponds to the tuning variable $c = 20$, [7]. The computation of the reference syllables are made using the same multitapers and parameter settings as for the computation of the analysis syllables in all cases.

In Figure 5, a colorplot of the errors in the time-frequency domain, $e^W_{j_r}(j_a)$, is seen for the LSP multitapers. The darkest blue color represent zero error where the analysis syllable is matched to the reference syllable corresponding to the same number, i.e., 1 belong to 1, 2 belong to 2, 8 belong to 8 and so on. However, it is also clearly seen that many of the following syllables of a reference syllable give small errors as well, e.g., syllable 3,4,5,6 and 7 also correspond fairly well to syllable number 2. We can then see the pattern of blue boxes, e.g, syllable 8 and 9 belong to reference syllable 8, and 10, 11 and 12 belong to 10 (although here the errors between 11 and 10 are larger). It can also be noted that the error is fairly large when comparing to other reference syllables (red and yellow colors). For this case and these multitapers, the classification is easily done without errors. The reason for the easy classification in the time-frequency domain is of course that modulation in especially frequency will cause great differences in the comparison. However, this is not necessarily

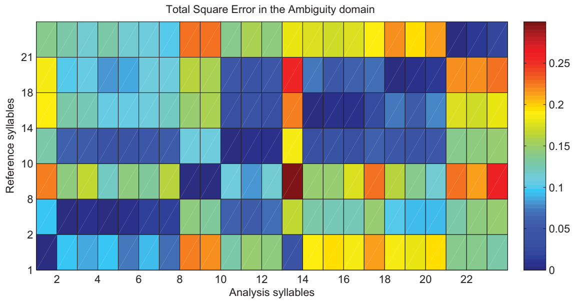

Figure 6: The total square error $e^{A}_{j_r}(j_a)$ of the ambiguity domain using the LSP multitapers. The blue color shows small error compared to the reference syllable (y-axis) where the red color shows large error.

Method	LSP	Thomson	Welch	Hann
Mean Value-W STD-W	.137 .0886	.140 .106	.146 .116	.374 .274
Mean Value-A STD-A	.0126 .00981	.0138 .00965	.0147 .0101	.0194 .0111
Mean Value-D STD -D	.0487 .0345	.0638 .0414	.0715 .0466	.167 .109

Table 1: The mean values and standard deviation of the total square errors $\min_{j_r} e^{z}_{j_r}(j_a)$ for all analysis syllables where $j_a \neq j_r$, for different methods and for different domains. Extension '-W' is the time-frequency domain, '-A', the ambiguity domain and '-D' the doppler domain.

wanted when comparing several different birds for finding similarities and differences in their singing. Using ambiguity domain analysis might give better tools for this. An example of the LSP multitapers applied for the ambiguity domain analysis is seen in Figure 6, which shows the same pattern as in Figure 5, for the darkest blue colors, and in this case, all three of 10,11 and 12 correspond in a more similar way to 10 (similar blue color). However, we also see that analysis syllables 10-12, 14-17, 18-20, all show such similarities that they all might be classified to reference syllable 10,14 and 18. Additionally, analysis syllables 2-7 are similar to reference syllable 10 and analysis syllables 10-12 are similar to reference syllable 2, which give us a total new view on what is a similarity and a difference. Further analysis with different birds is needed to see if the classification in this domain might be preferable for a certain material.

Finally, we compare the four different methods, single Hanning window (Hann), Welch method (Welch), Thomson multitapers (Thomson) and LSP multitapers (LSP), for robustness and analyze the mean value and standard deviation in the classification. The only method giving actual errors in the classification was the single Hanning window with 4 syllables erroneously classified in the doppler domain, and 1 syllable erroneously classified in the time-frequency and ambiguity domain respectively. None of the multitaper methods gave errors in the classification in any of the domains. This clearly shows the importance of robustness in the analysis methods where multitapers could be recommended. The mean value and standard deviation of the minimum total square errors $\min_{j_r} e^{z}_{j_r}(j_a)$ for all analysis syllables where $j_a \neq j_r$, are computed for the different methods and domains. The multitaper methods give similar results, clearly better than the single Hanning window. For all domains, the multitaper methods order with the smallest error for the LSP multitapers, followed by the Thomson multitapers and the Welch method.

6. CONCLUSIONS AND DISCUSSION

A novel method of syllable detection and classification of bird singing is presented. Also, different multitaper techniques are compared for time-frequency analysis of syllables and it is shown that multitaper techniques give a more robust classification. The locally stationary process multitapers gave the smallest error compared to the Thomson multitapers and Welch method. It is also proposed that the analysis and classification should be made in some other domain than the usual time-frequency representation (sonogram), e.g., the ambiguity or the doppler domain. Further analysis is however needed to be able to define the most appropriate domain. This might also depend on whether the evaluation is on the song properties of a same bird recorded at several occasions or a population study. For the extraction of classification features, the first left and right singular vectors value of the absolute value of the spectrum in respective domain is used. Other possibilities of classification features from the different spectra are left for further studies.

REFERENCES

[1] O. Tchernichovski, T. J. Lints, S. Deregnaucourt, A. Cimenser, and P. P. Mitra, “Studying the song development process. rationale and methods,” *Ann. NY Acad. Sci.*, vol. 1016, pp. 348–363, 2004.

[2] B. Boashash, “Theory of quadratic TFDs,” in *Time Frequency Signal Analysis and Processing; A Comprehensive Reference*, B. Boashash, Ed., chapter 3. Elsevier, 2003.

[3] D. J. Thomson, “Spectrum estimation and harmonic analysis,” *Proc. of the IEEE*, vol. 70, no. 9, pp. 1055–1096, Sept 1982.

[4] P. D. Welch, “The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms,” *IEEE Trans. on Audio Electroacoustics*, vol. AU-15, no. 2, pp. 70–73, June 1967.

[5] L. Cohen, *Time-Frequency Analysis*, Prentice-Hall, 1995.

[6] S. Aviyente and W. J. Williams, “Multitaper marginal time-frequency distributions,” *Signal Processing*, vol. 86, pp. 279–295, 2006.

[7] M. Hansson-Sandsten, “Optimal estimation of the time-varying spectrum of a class of locally stationary processes using hermite functions,” *EURASIP Journal on Advances in Signal Processing*, vol. 2011, 2011, Article ID 980805.

[8] R. A. Silverman, “Locally stationary random processes,” *IRE Trans. Inf. Theory*, vol. 3, pp. 182–187, 1957.