CLASSIFICATION OF BIRD SONG SYLLABLES USING SINGULAR VECTORS OF THE MULTITAPER SPECTROGRAM

Maria Hansson-Sandsten

Dept. of Mathematical Statistics, Lund University,
Box 118, SE-221 00 Lund, Sweden

ABSTRACT

Classification of song similarities and differences in one bird species is a subtle problem where the actual answer is more or less unknown. In this paper, the singular vectors when decomposing the multitaper spectrogram are proposed to be used as feature vectors for classification. The advantage is especially for signals consisting of several components which have stochastic variations in the amplitudes as well as the time- and frequency locations. The approach is evaluated and compared to other methods for simulated data and bird song syllables recorded from the great reed warbler. The results show that in classification where there are strong similar components in all the signals but where the structure of weaker components are differing between the classes, the singular vectors decomposing the multitaper spectrogram could be useful as features.

Index Terms— time-frequency, multitaper, spectrogram, SVD, bird song

1. INTRODUCTION

In the context of bird song, the aim is usually to classify bird *species*, e.g., [1–3]. The most popular features for classification relate to speech recognition, such as Linear Prediction Coefficients (LPC) and Mel-Frequency Cepstral Coefficients (MFCC). Other well known features are the spectrogram cross-correlation (SPCC), time- and frequency profiles or marginals, dynamic time warping (DTW) and fundamental frequency.

Classification of song similarities and differences in one *single bird species* is a more subtle problem where the actual answer is unknown. Similarities in songs from one year to another and similarities in songs from the same, contrary to distant populations, are still unexplored fields, 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 (GRW), which is the largest warbler species in Europe and a species with exceptional song capacity. A long-term study of a GRW population in Sweden is ongoing, see [4] and references therein, and one main aim is to understand the role of the song in an ecological and evolutionary context, [5]. It is notable that the recordings of GRW always are in natural environment and therefore are often noisy, e.g., from different wind conditions. It is noted in [1], that SPCC as well as DTW are sensitive to background noise. However, recently the usual noise sensitive spectrogram has been replaced by multitaper spectrograms, e.g., [6, 7] for better estimation of features.

In [8], the GRW song syllables are divided in the main classes of *whistles* and *rattles*, i.e., tonal sounds and several shorter components combined into a sound. The rattles are of a highly stochastic character in many different aspects, the number of components in two similar rattles might differ to some extent and the amplitudes of the different components are of stochastic character. The time-, and frequency locations of the components are also jittering when comparing two similar rattles. In this paper, the focus is on the rattle syllables and a non-stationary stochastic model is suggested for these sounds.

Singular value decomposition (SVD), as well as other techniques, e.g., non-negative matrix factorization (NMF), principal component analysis (PCA) and independent component anlysis (ICA) has been applied to time-frequency distributions with the aim to find especially noise reduced features or to extract dictionaries from a training data set for various applications. In most applications, the focus is on the singular values and the differences in them between classes. In [9, 10], the singular vectors (SV) of a positive time-frequency distribution are used to compute the time and spectral moments, indicating that the SVs are valuable in classification. In [11] we investigated the first pair of singular vectors of the multitaper spectrogram and used this as a feature comparing the inner products. The results indicated that the first SV pair of the multitaper spectrogram were not appropriate for classification of the GRW song syllables.

In this paper, a more thorough investigation is made of more than the first pair of SV and in the special case of multicomponent signals with stochastic variation in amplitudes as well as time- and frequency locations. An example is discussed, showing the advantage of using the SV and a scheme

Thanks to the Swedish Research Council and the eSSENCE academy for funding. Thanks also to the department of Biology, Lund University, for data collection.

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

for the inner product comparison is presented, which results in an optimal comparison between the feature vectors.

2. SIMILARITY MEASURE USING THE SINGULAR VECTORS OF THE SPECTROGRAM

The multitaper spectrogram is defined as

$$S_x(l,n) = \frac{1}{K}\sum_{k=1}^{K} |\sum_{n_1=0}^{N-1} x(n_1)h_k(n_1-n+M/2)e^{-i2\pi n_1 \frac{l}{2L}}|^2, \quad (1)$$

for $n = 0 \dots N-1$ and $l = 0 \dots L-1$ for time-discrete signals and frequencies. For $K = 1$, Eq. (1) is the usual windowed spectrogram using a length M unit energy window function. For $K > 1$, Eq. (1) is the multitaper spectrogram using a set of window functions, $h_k(n)$, $k = 1 \dots K$. In this paper the Hermite functions are used as windows as they have been shown to be the most localized in the time-frequency plane, [7].

The singular value decomposition (SVD) is a low-rank matrix approximation and a known technique for reducing noise in a data matrix. The decomposition of the $N \times L$ real valued spectrogram matrix $\mathbf{S}$ results in the representation

$$\mathbf{S} = \sum_{r=1}^{R} \sigma_r \mathbf{u}_r \mathbf{v}_r^T, \quad (2)$$

where superscript ^{T} denotes transpose, $\mathbf{u}_r, \mathbf{v}_r$ are the SV and $\sigma_1 \geq \dots \geq \sigma_R \geq 0$ the singular values. The unit-length vector $\mathbf{v}_1$, i.e. the first right singular vector, maximizes the norm $\|\mathbf{Sv}\|_2$ and can be seen as the vector which undergoes the maximum amplification under $\mathbf{S}$. Similarly $\mathbf{u}_1$ maximizes $\|\mathbf{S}^T\mathbf{u}\|_2$ and is the first approximation of the row-directions. Hence, in particular for low-rank $\mathbf{S}$, the vectors $\mathbf{u}_1$, $\mathbf{v}_1$ comprise most of the information in $\mathbf{S}$.

Recently, it has been pointed out that a usual non-negative factorization of a spectral matrix (image or time-frequency representation) is not unique and a non-negative factorization should be applied, [12]. With the approach presented in this paper, where the in-class data has more or less a similar appearance, this is not an issue.

The time profile, $\mathbf{t}_p = [t_p(0) \dots t_p(N-1)]^T$ with $t_p(n) = \frac{1}{L}\sum_{l=0}^{L-1} S_x(l,n)$ and frequency profile $\mathbf{f}_p = [f_p(0) \dots f_p(L-1)]^T$ with $f_p(l) = \frac{1}{N}\sum_{n=0}^{N-1} S_x(l,n)$, are very often used as a base for calculating features in classification, [2]. If the spectrogram matrix is a rank-one matrix, ($R = 1$), and the resulting time- and frequency profiles are normalized to unit energy, we can note that they in this special case are equal to the right SV $\mathbf{v}_1$ and left SV $\mathbf{u}_1$ respectively.

2.1. Pairwise similarity measure

As a measure of similarity of two syllables, the absolute value of the inner product in Euclidean space by $d(\cdot,\cdot) = |\langle \cdot,\cdot \rangle|$, of appropriate vectors is used. All vectors considered are normalized to one so the best possible similarity is $d(\cdot,\cdot) = 1$. Orthogonal vectors indicate difference and $d(\cdot,\cdot) = 0$ is the smallest possible value. The number of SV considered could be decided from the singular values and an appropriate threshold but in the paper we choose a fixed number Q of SVs from each matrix $\mathbf{S}$. Starting with the sets $\mathbf{U}^{S2} = [\mathbf{u}_1^{S2} \dots \mathbf{u}_Q^{S2}]$ and $\mathbf{V}^{S2} = [\mathbf{v}_1^{S2} \dots \mathbf{v}_Q^{S2}]$, we form all averages of the different combinations and find the vector pair in S2 that matches the one in S1,

$$q_2 = \arg\max(|(\mathbf{u}_{q1}^{S1})^T \mathbf{U}^{S2}| + |(\mathbf{v}_{q1}^{S1})_{S1}^T \mathbf{V}^{S2}|), \quad (3)$$

$$t(S1,S2)_{q_1} = (d(\mathbf{u}_{q1}^{S1}, \mathbf{u}_{q2}^{S2}) + d(\mathbf{v}_{q1}^{S1}, \mathbf{v}_{q2}^{S2}))/2 \quad (4)$$

for $q_1 = 1$. To avoid reusing a SV of S2, when repeating for other $q_1 = 2 \dots Q$, the matrices $\mathbf{U}^{S2}$ and $\mathbf{V}^{S2}$ are replaced by $\tilde{\mathbf{U}}^{S2}$ and $\tilde{\mathbf{V}}^{S2}$ where the columns numbered q_2 are removed before the next step.

A similar measure using the time- and frequency profiles is defined as

$$p(S1,S2) = (d(\mathbf{t}_p^{S1}, \mathbf{t}_p^{S2}) + d(\mathbf{f}_p^{S1}, \mathbf{f}_p^{S2}))/2. \quad (5)$$

2.2. An example

A simple synthetic syllable model for the rattles is proposed as

$$x(n) = \sum_{j=1}^{J} A_j \cos(2\pi F_j n + \phi) \cdot w_j(n - T_j) \quad n = 0 \dots N-1, \quad (6)$$

where $\phi \in R(0, 2\pi)$ and $w_j(n) = e^{-\sigma_j n^2}$ is a Gaussian window with σ_j chosen such that N_j^g values are above 0.01. 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 $T_j \in N(T_j^0, \sigma_{T_j})$.

To exemplify the approach of using SVD of the spectrogram, four syllables are simulated, see Figure 1, with $N = 800$, all $N_j^g = 128$ and the actual values of the parameters according to Table 1. We note the rather large differences of the amplitude parameters of syllables to be included in the same class. This could be seen as the natural variations of the syllables, one component is larger than the other or a component could even be missing in some cases, see real data example in Figure 5, where a last component is missing in half of the syllables in class 1. The features need to be robust against such variations.

First we study the two syllables S1-1 and S1-2, which have somewhat different component amplitudes but still obviously belong to the same class, starting with a low-frequency component followed by a high-frequency. We compare these syllables with S2-1, which should be of another class starting with a high-frequency followed by a low-frequency component. The corresponding spectrograms are computed using a

Class 1	**S1-2**	**S2-1**
$A_1/T_1/F_1$	0.80 / 200 / 0.05	1.0 / 200 / 0.05
$A_2/T_2/F_2$	0.90 / 400 / 0.15	0.60 / 400 / 0.15
Class 2	**S2-1**	**S2-2**
$A_1/T_1/F_1$	0.70 / 200 / 0.15	0.90 / 200 / 0.15
$A_2/T_2/F_2$	1.0 / 400 / 0.05	0.90 / 400 / 0.05
$A_3/T_3/F_3$		0.40 / 500 / 0.10
$A_4/T_4/F_4$	0.20 / 600 / 0.20	0.30 / 600 / 0.20

Table 1. The parameters of the four syllables in Figure 1

single Hanning window of length $M = 64$. The SVDs of the spectrograms give the left and right SV, where we study the first 4 pairs corresponding to the largest singular values. The left and right SV *outer* products $\mathbf{u}_1 \cdot \mathbf{v}_1^H$ are colored with blue for S1-1 and S1-2 in Figure 2 and $\mathbf{u}_2 \cdot \mathbf{v}_2^H$ are colored with red. We see that the positions are switched, caused by the fact that for S1-1 the high frequency component has the largest amplitude and for S1-2, the low-frequency component is largest. Using Eq. (4) we will find $t(S1\text{-}1, S1\text{-}2)_1 = 1.0$ and in the next step also $t(S1\text{-}1, S1\text{-}2)_2 = 1.0$. If we compare with the structure of the class 2 syllable S2-1, we can of course find e.g., $d(\mathbf{u}_1^{S1-1}, \mathbf{u}_2^{S2-1}) = 1.0$, also in this case, as the frequency locations are the same. However, the corresponding right SV inner products will be zero as these then are located at different time points. The result of Eq. (4) will be $t(S1\text{-}1, S2\text{-}1)_1 = 0.5$. The main point of this example is to show that if the correct combination of the SVs are made, it is a very robust tool for pairwise measures, where the variations of the component amplitudes are totally ignored, as the these values show up in the singular values.

In this example, using the time- and frequency profiles of Eq. (5) give $p(S1\text{-}1, S1\text{-}2) = 0.85$ and $p(S1\text{-}1, S2\text{-}1) = 0.94$, indicating more similarity between classes than in-classes. Using the time- or the frequency profiles individually do not change the result.

We might also have a member of class 2 as the one labeled S2-2 in Figure 1. Three components are similar to S2-1, see Table 1 but there is also an additional component between the middle and last one. We imaging that the syllable S2-2 comes from a more noisy measurement so white Gaussian noise is also added with an $SNR = 12.2$ dB, defined as mean power of the signal to the variance of the noise. Both these differences are possible, e.g., one or several components are often missing even if it is very clear that the structure of the syllable is the same. Two syllables to be compared might also come from different recordings and hence different amount of disturbances are possible. With all possible combinations we will find $t(S2\text{-}1, S2\text{-}2)_i \approx 0.999$, for $i = 1 \ldots 3$. The between class similarity will also in this case be $t(S1\text{-}1, S2\text{-}2)_i \approx 0.5$, for $i = 1 \ldots 2$.

Using the time- and frequency profiles give $p(S2\text{-}1, S2\text{-}2) = 0.94$ and $p(S1\text{-}1, S2\text{-}2) = 0.98$, again indicating higher similarity between classes than in-classes. From this example we see a number of great advantages, where the proposed SV-based similarity measure is very robust to amplitude differences as well as to noisy measurements and could be a better tool to use than the time- and frequency profiles for classification.

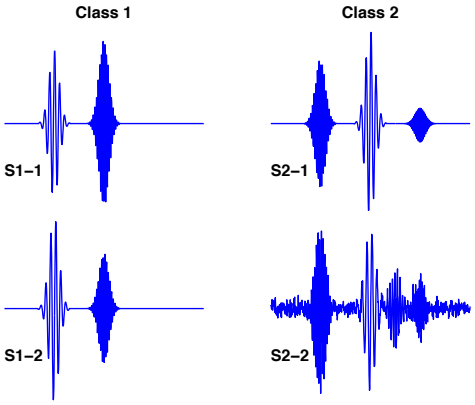

Fig. 1. Four examples of simulated syllables where S1-1 and S1-2 belong to class 1 and S2-1 and S2-2 belong to class 2.

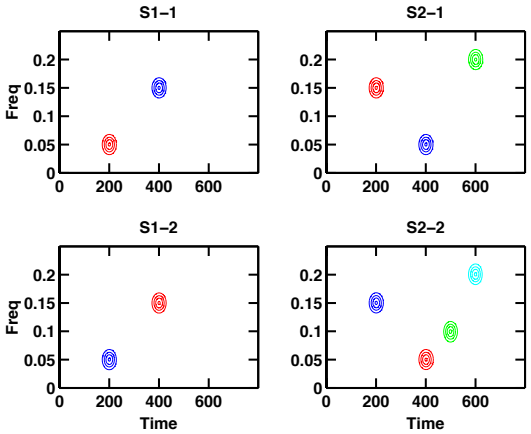

Fig. 2. The left and right singular vector pair outer products colored as $\mathbf{u}_1 \cdot \mathbf{v}_1^H$-blue, $\mathbf{u}_2 \cdot \mathbf{v}_2^H$-red, $\mathbf{u}_3 \cdot \mathbf{v}_3^H$-green, $\mathbf{u}_4 \cdot \mathbf{v}_4^H$-cyan

3. EVALUATION

The data model used is the one presented in Eq. (6), with $N = 1600$ and $N_j^g = 256$ for all values of j. A number of 20 syllables is generated in each class, where the stochastic amplitude, time- and frequency parameters are chosen according to Table 2. The variable σ is varied for different cases of the evaluation, increasing the jitter in amplitudes and time- and frequency locations of the components. Gaussian white noise were added to the signal with the SNR=13.9 dB. The multi-

Class 1	A_j^0, σ_{A_j}	F_j^0, σ_{F_j}	T_j^0, σ_{T_j}
$j=1$	$2, \sigma$	$0.3, 0.1\sigma$	$300, 300\sigma$
$j=2$	$0.8, \sigma$	$0.1, 0.1\sigma$	$800, 300\sigma$
$j=3$	$0.7, \sigma$	$0.4, 0.1\sigma$	$1300, 300\sigma$
Class 2	A_j^0, σ_{A_j}	F_j^0, σ_{F_j}	T_j^0, σ_{T_j}
$j=1$	$2, \sigma$	$0.3, 0.1\sigma$	$300, 300\sigma$
$j=2$	$0.7, \sigma$	$0.1, 0.1\sigma$	$1300, 300\sigma$
$j=3$	$0.8, \sigma$	$0.4, 0.1\sigma$	$800, 300\sigma$

Table 2. The parameters of the two classes in the evaluation, where σ is a variable in the simulations.

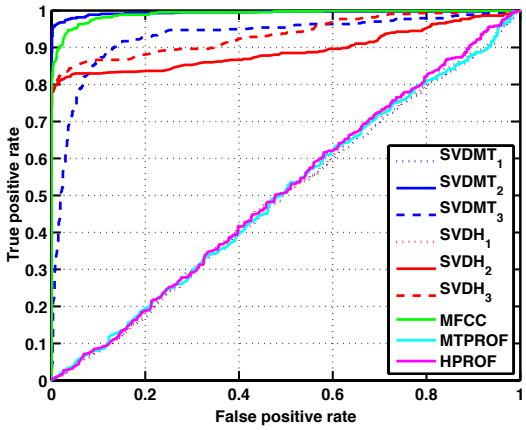

Fig. 3. The ROC-curves for the simulated binary-class data.

taper spectrogram (MT) is calculated using $K = 8$ Hermite functions, (from h_1 with $M = 100$ to h_8 with $M = 175$). The single Hanning window spectrogram (H) of length $M = 64$ is also applied. For these two different spectrograms, the SV are computed and $Q = 10$ SV pairs are ordered for the most similarity of two syllables finding $t(S1, S2)_1$ labeled as SVDMT_1 and SVDH_1 respectively. The next, with most similarity among the remaining SV, i.e., $t(S1, S2)_2$ is labeled SVDMT_2 and SVDH_2 and so on. This new approach is compared with the averaged profile measures in $p(S1, S2)$, (MTPROF and HPROF). We also include the MFCC method combined with pairwise DTW, using 8 cepstral coefficients, [3]. All unique pairwise combinations of the syllables are investigated and the sets of measures are divided into in-class (190 values) and between class (400 values) measures for further analysis.

The first evaluation is made for $\sigma = 0.1$ where the receiver operating characteristics (ROC) curves are calculated in Figure 3, with the correct in-class measures as the positive rate and it is clearly seen, as expected as the time- and frequency supports of the two classes are similar, that HPROF and MTPROF totally fails. We also see that SVDMT_1 and SVDH_1 also totally fails, caused by that the strongest component ($j = 1$) of the two classes are similar. The result

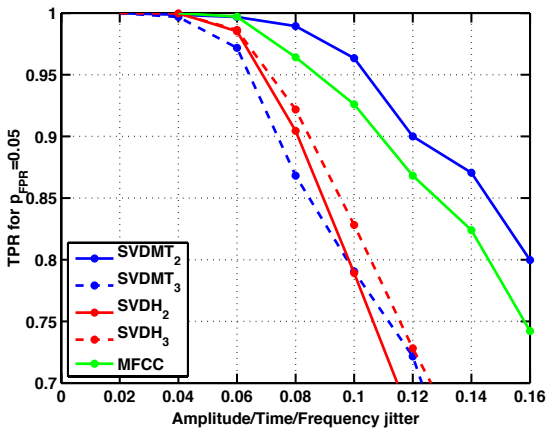

Fig. 4. The true positive rates allowing false positive rates of 5%.

from MFCC (green line) is much better, but the best result is given from SVDMT_2 (blue line very close to one on the y-axis) showing the robustness of the multitaper spectrogram. Next we simulate the syllables above for different values of σ ranging from zero to 0.16. The true positive rate were calculated for the false positive rate of 5% and the corresponding average results of 20 different simulations are depicted in Figure 4. The results show that the SVDMT_2 is superior to the other methods for all different values of σ. The methods that totally fails (HPROF, MTPROF, SVDMT_1 and SVDH_1) are not shown in the results.

4. REAL DATA

The method is evaluated on a small data set of two hand-sorted classes of syllables from one individual of the GRW. The syllables of a class are time-aligned using ordinary time-based correlation and are depicted in Figure 5. The variations in amplitudes are clearly seen as well as the missing last component in some of the syllables of class 1. There is also a variation of the number of small components coming just before the big component of the syllables of class 2. Below, the spectrograms of the first syllable in each class are depicted. The frequency contents are more or less the same as well as the time support, although there are clear differences in the distribution of components. Same parameters as used in the simulations are applied for the single Hanning window and the multitaper spectrograms. The choice of SV pairs were $Q = 10$. In Figure 6 the results from all different methods are shown. We see that this classification is not as difficult as the simulation case but still, there are differences in the performances of the methods. The best result is given for the SVDMT_2 followed by the SVDMT_3 and the MTPROF. These methods all give more than 90% correct classification allowing 5% false positives. This shows that these signals consist of more components and that more than one SV should be used

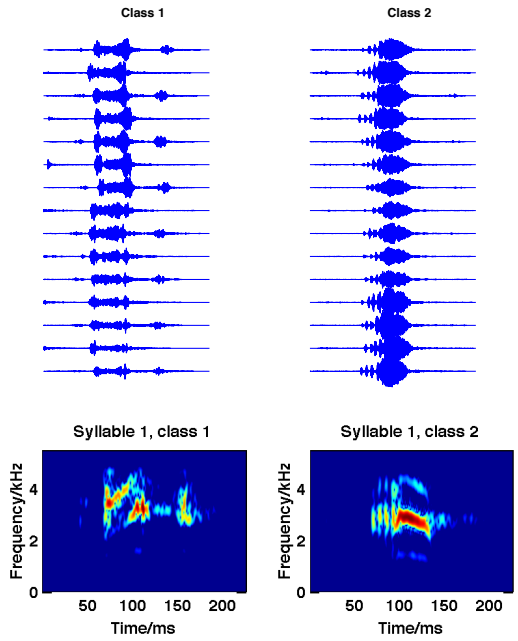

Fig. 5. The data in the two classes and two spectrogram examples of one syllable from each class.

for an appropriate classification. The MFCC and HPROF followed by the others are not appropriate for classification of these signals.

REFERENCES

[1] 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.

[2] F. Briggs, B. Lakshminarayanan, L. Neal, X. Z. Fern, and R. Raich, "Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach," *J. Acoust. Soc. Am.*, vol. 131, no. 6, pp. 4640–4650, 2000.

[3] 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.

[4] H. W. Lemke, M. Tarka, R. H. G. Klaassen, M. Åkesson, S. Bensch, D. Hasselquist, and B. Hansson, "Annual cycle and migration strategies of a trans-saharan migratory songbird: A geolocator study in the great reed warbler," *PLoS ONE*, vol. 8, no. 10, 2013.

[5] 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.

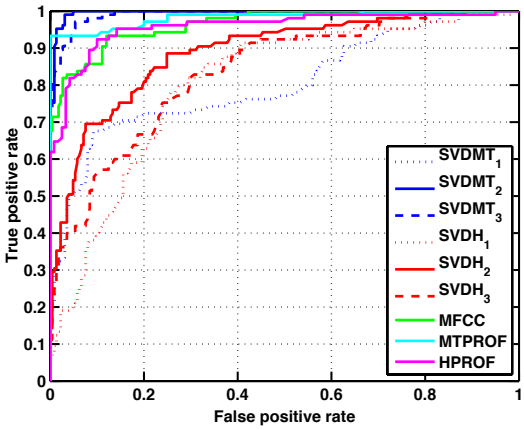

Fig. 6. The ROC-curves for the real data syllables of two classes.

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

[7] I. Daubechies, "Time-frequency localization operators: A geometric phase space approach," *IEEE Trans. on Information Theory*, vol. 34, no. 4, pp. 605–612, 1988.

[8] E. Wegrzyn, K. Leniowski, and T. S. Osiejuk, "Whistle duration and consistency reflect philopatry and harem size in great reed warblers," *Animal Behaviour*, vol. 79, no. 6, pp. 1363–1372, 2010.

[9] D. Groutage and D. Bennink, "Feature sets for non-stationary signals derive from moments of the singular value decomposition of cohen-posch (positive time-frequency) distributions," *IEEE Trans. on Signal Processing*, vol. 48, no. 5, pp. 1498–1503, May 2000.

[10] B. Ghoraani, "Selected topics on time-frequency matrix decomposition analysis," *Journal of Pattern Recognition and Intelligent Systems*, vol. 1, no. 3, pp. 64–78, 2013.

[11] M. Hansson-Sandsten, M. Tarka, J. Caissy-Martineau, B. Hansson, and D. Hasselquist, "A SVD-based classification of bird singing in different time-frequency domains using multitapers," in *European Signal Processing Conference (EUSIPCO)*, Barcelona, Spain, 2011.

[12] J. Gruninger and H. Dothe, "Boundary constraints for singular value decomposition of spectral data," *Proc. of SPIE*, vol. 8892, 2013, doi:10.1117/12.2029250.