Indonesian Journal of Electrical Engineering and Computer Science
Vol. 14, No. 1, April 2019, pp. 59~64
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v14.i1.pp59-64

Improving spectrogram correlation filters with time-frequency reassignment for bio-acoustic signal classification

Salina Abdul Samad, Aqilah Baseri Huddin
Centre for Integrated Systems Engineering and Advanced Technologies (INTEGRA),
Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia, Malaysia

Article Info

Article history:

Received Jun 21, 2018
Revised Dec 12, 2018
Accepted Jan 17, 2019

Keywords:

Bio-acoustic signal
Classification
Correlation filter
Spectrogram
Time-frequency reassignment

ABSTRACT

Spectrogram features have been used to automatically classify animals based on their vocalization. Usually features are extracted and used as inputs to classifiers to distinguish between species. In this paper, a classifier based on Correlation Filters (CFs) is employed where the input features are the spectrogram image themselves. Spectrogram parameters are carefully selected based on the target dataset in order to obtain clear distinguishing images termed as call-prints. An even better representations of the call-prints are obtained using spectrogram Time-Frequency (TF) reassignment. To demonstrate the application of the proposed technique, two species of frogs are classified based on their vocalization spectrograms where for each species, a correlation filter template is constructed from multiple call-prints using the Maximum Margin Correlation Filter (MMCF). The improved accuracy rates obtained with TF reassignment demonstrates that this is a viable method for bio-acoustic signal classification.

Corresponding Author:

Salina Abdul Samad
Centre for Integrated Systems Engineering and Advanced Technologies (INTEGRA),
Faculty of Engineering and Built Environment,
Universiti Kebangsaan Malaysia,
43600 UKM Bangi, Selangor, Malaysia.
Email: salinasamad@ukm.edu.my

1. INTRODUCTION

Many animals are routinely identified by to their vocalization especially in cases where their images are not available. This is usually done manually by experts from field recordings. This endeavor has wide ranging applications in the field of natural habitat conservation. For example, the frog population is used as one of the bio-indicators in assessing the health of habitats such as wetlands and floodplains [1]. This has sparked interest in using automation to identify and classify animals from their vocalizations. This area of bio-acoustics signal analysis has been mostly concentrated on using techniques similar to those used for processing speech signals [2], [3]. Signal segmentation of bio-acoustic sounds is routinely performed in order to isolate syllables [4], [5]. The next step usually involved extracting relevant features, such as the popular Mel Frequency Cepstrum Coefficients (MFCC) [6], [7], or other simpler features based on parameters such as sound or call duration, maximum power, and maximum and minimum frequencies [8]. Akin to speech processing, once the features are identified, they are used as inputs to classifiers such as Support Vector Machines [9], Nearest Neighbors [10], Neural Networks [11], [12] and many others [13-15].

Another approach to process animal calls are based on spectrograms. Spectrograms are visual representations of audio signals obtained using the Short-Time Fourier Transform (STFT). Image processing techniques commonly used in many applications [16], have been applied to spectrograms to automatically analyzed animal calls [17]. Researchers in animal calls have extracted features from the spectrograms such

as local peaks [18] and ridges [19]. These features, or their derivatives, are then used as inputs to classifiers similar to those described above for speech processing.

This paper presents a different approach to animal call classification. It investigates the possibility of representing animal call spectrograms as call-prints similar to fingerprints in humans. As call-prints are images, a classification technique based on Correlation Filters (CFs) are applied to the images. There are many types of CFs that have been used for image classification. They have been designed to exhibit attractive properties such as noise robustness, shift-invariance, distortion tolerance and gradual degradation. A such, they have been applied to image processing applications such as biometric classification, pedestrian detection, object detection and tracking [20], [21]. In these applications, a template (also known as filter) is carefully designed from the training images. A query image is then cross-correlated with the template to produce the output, where the operation is performed in the frequency domain in order to take advantage of the efficiencies of the Fast Fourier Transform (FFT) algorithm.

The Maximum Margin Correlation Filter (MMCF) is considered in this paper because it has been shown to perform better than other well-known types of CF such as Optimal Trade-off Synthetic Discriminant Function (OTSDF), Unconstrained Optimal Trade-off Synthetic Discriminant Function (UOTSDF), Average of Synthetic Exact Filter (ASEF) and Minimum Output Sum of Squared Errors (MOSSE). The MMCF has been designed to provide the advantage of providing not only good classification but also of localization as demonstrated successfully in tasks such as vehicle recognition, eye localization and face classification [22], [23].

In this paper, the MMCF is applied to bio-acoustic signal spectrograms as the images to construct the templates. Several training images are used to synthesize a filter template. In order to obtain distinctive animal call-prints, the spectrograms have to be carefully constructed by centering the call in an image frame and by selecting parameters that highlight salient features of the calls. An even better representation of the call-print is obtained using time-frequency reassignment. To demonstrate the viability of this technique, it is applied to a two-class task, classifying two different frog species based on their calls.

2. RESEARCH METHOD

The proposed technique consists of three main parts: constructing call-prints using spectrograms, constructing MMCF templates using multiple call-prints and classifying animal vocalization using correlation plane parameters.

2.1. Constructing call-prints

2.1.1. Spectrogram

It is important to obtain a good representation of an animal call in terms of its spectrogram representation. This call-print is dependent on the parameters used for the spectrogram, which in turn is dependent on the dataset. In this case, the recordings of the two species of frogs sampled at 44.1 kHz, are segmented into individual calls of 800 ms length using a sound editing tool. Each segment is then filtered with a high pass filter with a cut-off frequency of 250 Hz in order to eliminate the environmental noise. A centering of the peak amplitude at 400 ms is performed before applying the STFT in order to obtain the spectrogram.

Framing are performed on the calls with a chosen frame length of 256 with a 75 percent overlap. These parameters are obtained after several trials with the objective of obtaining visually clear call-prints. Windowing is applied using the Gaussian window chosen due to its superior performance in eliminating energy leakage, even though the computation is more intensive compared to other windows [23]. The Gaussian window is described by

$$h(n) = e^{-\frac{1}{2}\left(\frac{2.5n}{N/2}\right)^2},\ 0 \le |n| \le \frac{N}{2} \quad (1)$$

where N is the window length.

Each windowed frame is then transformed from time domain into the frequency-domain by the STFT to construct the spectrogram of the magnitude spectrum, also termed here as the call-prints.

2.1.2. Time-frequency reassignment

Time-frequency (TF) reassignment is a technique used to overcome the shortcomings of spectrograms due to the unfortunate trade-off between resolution in frequency and time [24]. As such, TF reassignment is applied to obtain an even better representation of the call-prints. TF reassignment uses the information from the phase spectrum to sharpen the amplitude spectrum. It is able to locate impulses, linear

chirps, and simple sinusoids at the actual time or frequency with a higher resolution than the inherent STFT spectrograms. It has been shown that TF reassignment is equivalent to moving energy up the local gradient of intensity of the spectrogram for Gaussian windows [25].

Utilizing the modified moving window method [24], consider a set of coefficients $\varepsilon(t,\omega)$ obtained by decomposing a time domain signal $x(t)$ based on a set of elementary signals $h_\omega(t)$ where

$$h_\omega = h(t)e^{j\omega t} \tag{2}$$

with h(t) being a lowpass kernel function. The coefficients of this decomposition are defined as

$$\epsilon(t,\omega) = \int x(\tau)h(t-\tau)e^{-j\omega(\tau-t)}d\tau \tag{3}$$

resulting in

$$\epsilon(t,\omega) = e^{j\omega t}X(t,\omega) = X_t(\omega) \quad = M_t(\omega)e^{j\emptyset_\tau(\omega)} \tag{4}$$

Here, $M_t(\omega)$ is the magnitude and $\varphi(\omega)$ is the phase of $X_t(\omega)$, the Fourier transform of the signal $x(t)$ shifted by time t and windowed by $h(t)$.

For signals exhibiting slow time variation compared to phase variation, the maximum contribution to the reconstruction integral comes from the vicinity of the point t,ω satisfying the phase stationary condition

$$\frac{\partial}{\partial\omega}[\emptyset_\tau(\omega) - \omega\tau + \omega\tau] = 0 \tag{5a}$$

$$\frac{\partial}{\partial\tau}[\emptyset_\tau(\omega) - \omega\tau + \omega\tau] = 0 \tag{5b}$$

Or equivalently around the point $\hat{t},\hat{\omega}$ defined by

$$\hat{t}(\tau,\omega) = -\frac{\partial\emptyset(\tau,\omega)}{\partial\omega} \tag{6a}$$

$$\hat{\omega}(\tau,\omega) = \omega + \frac{\partial\emptyset(\tau,\omega)}{\partial\tau} \tag{6b}$$

This method of reassignment changes the point of attribution of $\varepsilon(t,\omega)$ to this point of maximum contribution $\hat{t}(t,\omega),\hat{\omega}(t,\omega)$ rather than to the point of t,ω at which it was originally computed. The computed time-frequency coordinates $\hat{t}(t,\omega),\hat{\omega}(t,\omega)$ are equal to the local group delay and the local instantaneous frequency, respectively. These quantities are normally ignored when constructing the spectrogram which only considers the magnitude.

2.2. Constructing templates

The bio-acoustic signal spectrograms are used to construct the MMCF template. Several training images are used to synthesize a filter template. Each image is of size 512x512 and the template is constructed from multiple call-prints from the training set. A separate test set is used for the cross-correlation process with the template filter in order to determine whether the test image is from the true or false class. In this process, the MMCF optimizes a criterion to produce a desired correlation output plane by a trade-off matrix maximizing the margin criterion similar to Support Vector Machine (SVM), while minimizing the localization criterion expressed as the mean square error. As with other CFs, the MMCF can be expressed in a closed form solution [20]. As such, the optimization of the MMCF template can be described as

$$\boldsymbol{H}_{MMCF} = \widetilde{\boldsymbol{T}}^{-1}\frac{1}{L}\left(\sum_{i=1}^{L}\widetilde{\boldsymbol{X}}_i\boldsymbol{g}_i\right) + \widetilde{\boldsymbol{T}}^{-1}\boldsymbol{A}\widetilde{\boldsymbol{Y}}\boldsymbol{a} \tag{7}$$

where $\widetilde{\boldsymbol{T}}$ is the trade-off matrix, $\widetilde{\boldsymbol{X}}_i$is a d x d diagonal matrix form of the ith training image in the frequency domain with vector $\boldsymbol{x}_i$ along its diagonal, $\boldsymbol{g}_i$ is the 2D vector representation of the expected correlation output for the ith training image, $\boldsymbol{A}$ is a d x L matrix whose columns are formed by L training image vectors $\boldsymbol{x}_i$, $\widetilde{\boldsymbol{Y}}$ is a diagonal matrix with class label (1 for true class, 0 for false class) along its diagonal, while the vector $\boldsymbol{a}$ is evaluated from the sequential minimum optimization technique.

2.3. Classification

The template matching process for the input test image *S(x,y)* and a correlation filter template *H(u,v)* is given by

$$c(x,y) = IFFT\{FFT\big(S(x,y)\big) * H^{+}(u,v)\} \quad (8)$$

The test image is first converted to the frequency domain and then reshaped to be in the form of a vector. It is then convolved with the conjugate of the MMCF, or equivalently, cross correlating it with the MMCF. Transformation of the output to the spatial domain is required in order to obtain the correlation plane.

If the test image belongs to the same class as the designed filter, the resulting correlation plane produces a sharp peak at the origin while the values everywhere else are close to zero. To measure the sharpness of the peak, the Peak-to-Sidelobe ratio (PSR) is used, where

$$PSR = \frac{peak-mean}{standard\ deviation} \quad (9)$$

The peak is the largest value of the test image obtained from the correlation output. The standard deviation and mean are calculated from a sidelobe region excluding a central mask [17]. To classify the frogs into the correct class, threshold value for each class is determined from the PSR values obtained with the cross-correlation process using the dataset. Then, the cross correlation process is performed using the test set. If an image has a PSR value that is greater than the threshold for the tested class, it is classified as its true class, otherwise it is classified as false.

3. RESULTS AND ANALYSIS

To demonstrate the viability of using the proposed technique to classify call-prints, two species of frogs commonly found in Malaysia were considered. Recordings were obtained for common grass frogs (*F. limnocharis Boie*) and mangrove frogs (*F. cancrivora Gravenhorst*). The calls were subsequently processed as described in section 2.1 to obtain the call-prints. For each species, 30 call-prints were obtained and divided equally into the training and testing sets. The templates for each class were constructed as described in section 2.2, using 5, 10 and 15 call-prints from the test-set. From the cross-correlation process of the test set and the templates as described in section 2.3, the accuracy rate is calculated defined as the ratio of correct classification to total number of test inputs. The results are tabulated in Table 1 to 3 for different number of call-prints per template, with and without TF reassignment.

Table 1. Accuracy Rate with 5 Call-Prints per Template

Species	Without TF Reassignment (%)	With TF Reassignment (%)
F. limnocharis Boie	25.3	34.7
F. cancrivora Gravenhorst	15.7	21.1
Average	20.5	27.9

Table 2. Accuracy Rate with 10 Call-Prints per Template

Species	Without TF Reassignment (%)	With TF Reassignment (%)
F. limnocharis Boie	45.9	60.7
F. cancrivora Gravenhorst	60.5	73.1
Average	53.2	66.5

Table 3. Accuracy Rate with 15 Call-Prints per Template

Species	Without TF Reassignment (%)	With TF Reassignment (%)
F. limnocharis Boie	67.5	79.0
F. cancrivora Gravenhorst	73.1	86.8
Average	70.3	82.9

The Tables show that in general, for cases with and without TF reassignment, the accuracy rate increases as the number of call-prints per template increases. The results also demonstrate that using TF reassignment increases the accuracy rate by more than 10 percent for the cases of 10 and 15 call-prints per template suggesting that this is viable technique that can be used to distinguish between animal species based on their vocalization spectrogram.

4. CONCLUSION

This paper has shown that bio-acoustic signals may be classified using MMCFs when the signals are converted to spectrograms in order to obtain the call-prints. Call-prints has been shown to be viable image inputs to MMCFs for frog classification based on their vocalization. However, careful selection of the spectrogram parameters is required in order to produce clear and distinguishing call-prints. Multiple call-prints are used to construct the MMCF template representing the species. The results of classifying two frog species showed that the accuracy rate increases as the number of call-prints per template increases. Furthermore, applying TF reassignment to the spectrograms increases the accuracy rate overall and by more than 10 percent for 10 and 15 call-prints per template

ACKNOWLEDGEMENTS

The authors wish to acknowledge the contribution of the Ministry of Education Malaysia through the funding of project FRGS/1/2016/TK04/UKM/01/1

REFERENCES

[1] J.P. Gibbs, S. Rouhani, and L. Shams. "Frog and Toad Habitat Occupancy across a Polychlorinated Biphenyl (PCB) Contamination Gradient," *Journal of Herpetology*, vol. 51, no. 2, pp. 209-14, 2017.
[2] S. N. Endah, S. Adhy, and S. Sutikno, "Comparison of Feature Extraction Mel Frequency Cepstral Coefficients and Linear Predictive Coding in Automatic Speech Recognition for Indonesian," *TELKOMNIKA*, vol. 15, no. 1, pp. 292–298, 2017
[3] M. F. Alghifari, T. S. Gunawan, and M. Kartiwi, "Speech Emotion Recognition using Deep Feedforward Neural Network," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 10, no. 2, pp. 554–561, 2018
[4] C.J. Huang, Y.J. Yang, D.X. Yang, and Y.J Chen, "Frog Classification using Machine Learning Techniques," *Expert Systems with Applications*, vol. 36, vol. 2, pp. 3737-3743, 2009
[5] H. Jaafar, and D.A Ramli, "Automatic Syllables Segmentation for Frog Identification System," *Proceedings of IEEE 9th International Colloquium on Signal processing and Its Applications (CSPA),* pp. 224-228, 2013.
[6] C.H. Lee, C.H. Chou, C.C. Han, and R.Z Huang, "Automatic Recognition of Animal Vocalizations using Averaged MFCC and Linear Discriminant Analysis," *Pattern Recognition Letters*, vol. 27, no. 2, pp. 93-101, 2006.
[7] B. Gingras, and W.T Fitch, "A Three-Parameter Model for Classifying Anurans into Four Genera Based on Advertisement Calls," *The Journal of the Acoustical Society of America*, vol. 133, no. 1, pp. 547-559, 2013
[8] M.A. Acevedo, C.J. Corrada-Bravo, H. Corrada-Bravo, L.J. Villanueva-Rivera, and T.M. Aide, "Automated Classification of Bird and Amphibian Calls using Machine Learning: A Comparison of Methods," *Ecological Informatics*, vol.4, no. 4, pp. 206-214, 2009
[9] J.J. Noda, C.M. Travieso, and D. Sánchez-Rodríguez, "Methodology for Automatic Bioacoustic Classification of Anurans Based on Feature Fusion," *Expert Systems wiith Applications*, vol. 50, pp.100-106. 2016
[10] G. Vaca-Castaño, and D. Rodriguez, "Using Syllabic Mel Cepstrum Features and K-Nearest Neighbors to Identify Anurans and Birds Species," *Proceedings of the IEEE Workshop on Signal Processing Systems (SIPS),* pp. 466-471, 2010.
[11] C.J. Huang, Y.J. Chen, H.M. Chen, J.J. Jian, S.C. Tseng, Y.J. Yang, and P.A Hsu, "Intelligent Feature Extraction and Classification of Anuran Vocalizations," *Applied Soft Computing*, vol. 19, pp. 1-7, 2014.
[12] J. Colonna, T. Peet, C.A. Ferreira, A.M. Jorge, E.F. Gomes, and J. Gama, "Automatic Classification of Anuran Sounds using Convolutional Neural Networks," *Proceedings of the Ninth International C* Conference on Computer Science & Software Engineering*, pp. 73-78, 2016.
[13] N.C. Han, S.V. Muniandy, and J. Dayou, "Acoustic Classification of Australian Anurans Based on Hybrid Spectral-Entropy Approach," *Applied Acoustics*, vol. 72, no. 9, pp. 639-645, 2011.
[14] B. Gingras, and W.T. Fitch, "A Three-Parameter Model for Classifying Anurans into Four Genera Based on Advertisement Calls," *The Journal of the Acoustical Society of America*, vol. 133, no. 1, pp. 547-559, 2013.
[15] W.P. Chen, S.S. Chen, C.C. Lin, Y.Z. Chen, and Lin W.C., "Automatic Recognition of Frog Calls Using a Multi-Stage Average Spectrum," *Computers & Mathematics with Applications*, vol. 64, no. 5, pp. 1270-1281, 2012.
[16] K. Arun Sai, and K. Ravi, "An Efficient Filtering Technique for Denoising Colour Images," *International Journal of Electrical and Computer Engineering (IJECE),* vol. 8, no. 5, pp. 3604-3608, 2018.
[17] J. Dennis, H.D. Tran, and H. Li, "Spectrogram Image Feature for Sound Event Classification in Mismatched Conditions,", *IEEE Signal Processing Letters*, vol. 18, no. 2, pp.130-133, 2011.

[18] G. Grigg, A. Taylor, H. Mc Callum, and G. Watson, "Monitoring Frog Communities: An Application of Machine Learning," *Proceedings of Eighth Innovative Applications of Artificial Intelligence Conference*, pp. 1564-1569, 1996.
[19] J. Xie, M. Towsey, J. Zhang, X. Dong, and P Roe, "Application of Image Processing Techniques for Frog Call Classification," *Proceedings of IEEE International Conference on Image Processing (ICIP)*, pp. 4190-4194, 2015.
[20] Q. Wang, A. Alfalou, and C. Brosseau, "New Perspectives in Face Correlation Research: A Tutorial, "*Advances in Optics and Photonics*," vol. 9, no. 1, pp. 1-78, 2017.
[21] R.A. Kerekes, and B.V. Kumar, "Selecting A Composite Correlation Filter Design: A Survey and Comparative Study," *Optical Engineering*, vol. 47, no. 6, pp. 067202, 2008.
[22] A. Rodriguez, V.N. Boddeti, B.V. Kumar, and A. Mahalanobis, "Maximum Margin Correlation Filter: A New Approach for Localization and Classification," *IEEE Transactions on Image Processing*, vol. 22, no. 2, pp. 631-643, 2013.
[23] A.F. Rodriguez-Perez, "Maximum Margin Correlation Filters," *Doctoral Dissertation*, Carnegie Mellon University, 2012.
[24] P. Flandrin, F. Auger, and E Chassande-Mottin, "Time-Frequency Reassignment: From Principles to Algorithms," *Applications in Time-Frequency Signal Processing*, vol. 5, on. 102, pp.179-203, 2003.
[25] E. Sejdić, Djurović I., and J. Jiang, "Time-Frequency Feature Representation using Energy Concentration: An Overview of Recent Advances," *Digital Signal Processing*, vol. 19, no.1, pp.153-183, 2009.

BIOGRAPHIES OF AUTHORS

Salina Abdul Samad obtained a Bachelor of Science in Electrical Engineering from the University of Tennessee, USA in 1986 and a Ph.D. from the University of Nottingham, UK in 1995. She is a professor of Signal Processing at the Centre for Integrated Systems Engineering and Advanced Technologies (INTEGRA), Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia (UKM), where she heads the Signal Processing research group.

Aqilah Baseri Huddin received her BEng (Hons) in Electrical and Electronics Engineering and her PhD in Electrical and Electronics Engineering from the University of Adelaide, Australia in 2007 and 2015, respectively. She is currently a senior lecturer at the Centre for Integrated Systems Engineering and Advanced Technologies (INTEGRA), Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia (UKM). Her research interests are mainly in the field of image processing and artificial intelligence