Circuits, Systems, and Signal Processing (2022) 41:2669–2680
https://doi.org/10.1007/s00034-021-01896-2

Bird Call Classification Using DNN-Based Acoustic Modelling

Rajeev Rajan[1,2] · Jisna Johnson[1,2] · Noumida Abdul Kareem[1,2]

Received: 18 April 2021 / Revised: 21 October 2021 / Accepted: 23 October 2021 /
Published online: 3 January 2022

Abstract

Bird call recognition using deep neural network-hidden Markov model (DNN-HMM)-based transcription is proposed. The work is an attempt to adapt the human speech recognition framework for bird call classification through transcription approach. Initially, the phone transcriptions are generated using CMU-Sphinx, and lexicons are modified using group delay-based segmentation. Later, bird call transcription is implemented using hybrid DNN-HMM framework through DNN-based acoustic modelling. During the DNN-based acoustic modelling, mel-frequency cepstral coefficient features (MFCCs) are computed and experimented with monophone models, triphone models, followed by linear discriminative analysis and maximum likelihood linear transform. The transcribed phonemes are corrected using context-based rules in the final phase. The proposed approach is evaluated on a dataset that consists of ten species with 563 audio tracks. The hybrid DNN-HMM approach outperforms the convolutional neural network and long short-term memory framework with an accuracy of 94.46%.

Keywords Hidden Markov model · Gaussian mixture model · Deep neural network · Convolutional neural network

1 Introduction

In the last few decades, considerable research efforts have been devoted to the automatic analysis of speech. However, research in automatic analysis of vocalizations

✉ Rajeev Rajan
rajeev@cet.ac.in

Jisna Johnson
jisnajohnson9@gmail.com

Noumida Abdul Kareem
noumidaa@gmail.com

[1] Department of Electronics and Communication Engineering, College of Engineering, Trivandrum, Thiruvananthapuram, India

[2] APJ Abdul Kalam Technological University, Thiruvananthapuram, Kerala, India

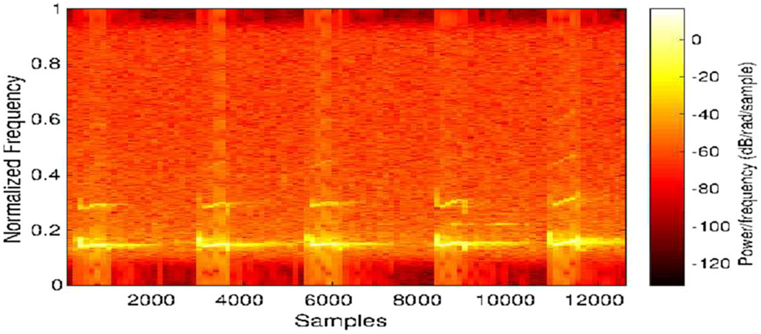

Fig. 1 Spectrogram of bird's vocalization of owl

from animals and birds has intensified only recently. Bird vocalization plays a vital role in communication between species, and it includes both bird calls and bird songs. In ornithology, songs are treated as relatively complex vocalizations as compared to the calls, which are considered as simple vocalizations. [1] A bird may listen to other birds and classify them as neighbour or stranger, kin or non-kin. Birds may also sing to attract a mate or for territory defence. The vocalization pattern of the bird owl can be noticed from the spectrogram in Fig. 1. Ornithologists are interested in finding out whether a particular perhaps rare species has appeared in a given region or to find out how many different bird species occur in a given region. Traditional field techniques to track and identify different bird species have required much human effort. Acoustic signal-based bird monitoring is an effective approach as most birds use vocalizations as a primary communication method [10]. Besides, ecological and behavioural studies can also be benefited from the automatic detection of bird species from raw audio field recordings.

Global biodiversity information facility (GBIF), [2] which creates biological multimedia databases, also focuses on automatic classification and identification of species from field recordings. Moreover, the automation gained momentum due to the advancement of bio-acoustic signal processing and pattern recognition techniques [7,23]. Several techniques of speech and audio processing can also be applied well to bird calls [6,24,28]. We explore the effectiveness of fully connected DNNs for the transcription of bird calls in similar lines as they are used for speech and language identification tasks [23].

Numerous attempts can be seen in the literature for bird species classification and transcription through acoustic cues [19,26,27]. In [2], the HTS (HMM-based speech synthesis system) framework is proposed to model and synthesize bird songs. Bird species recognition through unsupervised modelling of individual bird syllables and duration modelling is explored in [9]. An iterative maximum likelihood procedure is introduced in the above work, to train the individual HMMs for syllables of each species. An HMM-based detector with a general model trained from all the syllables is used as a baseline system in [7]. In an improved system, syllable patterns are first

[1] https://en.wikipedia.org/wiki/Bird-vocalization.

[2] www.gbif.org.

inferred from similar syllables observed in the recordings. HMMs of the inferred syllable patterns are then trained to allow finer acoustic modelling of the syllables. Experimental results show that the use of individual element HMMs of bird species improved the identification accuracy significantly in comparison to the single HMM [10]. Dynamic kernel-based SVMs and DNNs for classification of bird calls can be refereed in [3]. Acoustic detection of avian sounds in terms of syllables is also used for bird recognition [19].

Deep convolutional neural networks have also been applied efficiently in bird species recognition. Splitting the audio recording into temporal segments, passing it to the CNN and receiving output for each segment is experimented in [15]. A data-efficient bird call classification is implemented using CNN transfer learning approach in [5]. Conventional hidden Markov models with a probability density function at each state are employed for bird species recognition in [11] and [9]. In the proposed task, we attempted hybrid DNN-HMM transcription approach for bird species classification from their calls.

Section 2 describes the proposed system. Evaluation of the system is discussed in Sect. 3, followed by analysis of the results in Sect. 4. Finally, the conclusion is drawn in Sect. 5.

2 Proposed System

Block diagram of the proposed scheme is shown in Fig. 2. Phoneme transcriptions are obtained from CMUSphinx-framework in the first phase. Later, lexicon is modelled using phonetic transcriptions and boundary information obtained from the group delay-based segmentation. The subsequent DNN-based acoustic modelling uses this lexicon for its transcription. During the DNN-based acoustic modelling, MFCCs are computed and experimented with monophone and triphone models, followed by linear discriminative analysis (LDA) and decorrelation using maximum likelihood linear transform (MLLT). Finally, in the hybrid DNN-HMM framework, the HMM topology is generated based on GMM-HMM model and transcription is done. The steps in the

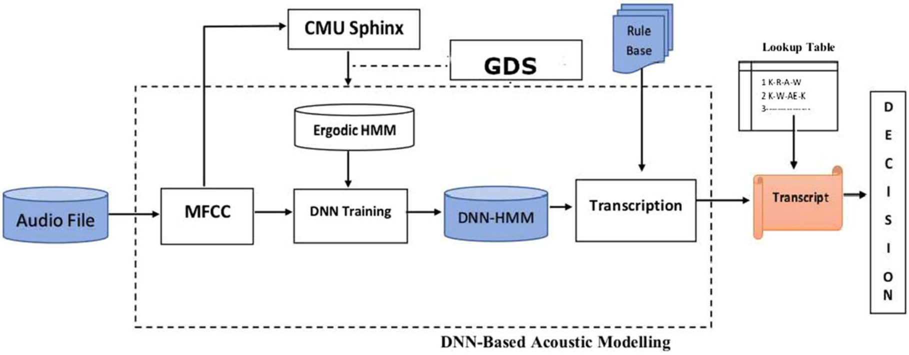

Fig. 2 Block diagram of DNN-HMM framework for the bird classification system

Table 1 Initial phone level transcription, TR and TT denote train and test files, respectively

Sl.No	Bird Species	Lexicon	# TR	# TT
1	House crow	K-R-W-A-W	14	50
2	Mallard duck	KW-AE-K	30	30
3	Asian koel	OW-UW-HH	23	37
4	Eurasian owl	H-U-OO-T	28	40
5	House sparrow	T-Z-HH-IY	14	35
6	Blue jay	T-UW-UW	33	20
7	Red-wattled lapwing	T-OW-T-OW-T-UW	30	20
8	Grey go-away	TH-OW-R-EY	20	27
9	Indian peafowl	EY-OW-HH	23	36
10	Western wood pewee	P-EE-W-EE	23	30
	Total		238	325

phoneme transcription and DNN-based acoustic modelling are explained in following subsections. The transcript obtained is corrected through context-based rules using a look-up table, which includes reference lexicon and bird-tag name. The final transcript is checked for matching with the lookup table for the bird name. The performance of the proposed scheme is compared with that of feature-classifier strategies such as MFCC-DNN, spectrogram-CNN and MFCC-LSTM frameworks.

2.1 Phonetic Transcriptions

CMUSphinx is used for generating a transcription in the first phase. In CMUSphinx, the adaptation of the English model to bird call vocalization is performed by creating a phone-set map and English phone set dictionary. The fit between the bird vocalization and the English language model can potentially be improved by the adaptation of English model using bird calls in the corpus. The efficacy of the model is improved by adaptation, and this model is used for transcription. CMUSphinx provides phonetic transcriptions which is used for building lexicons in the subsequent DNN-based acoustic modelling. The initial lexicon prepared using phone level transcription is shown in Table 1.

Initial lexicons are updated by incorporating suprasegmental domain knowledge to aid better transcription next phase. A simple candidate for segmenting acoustic signal into syllable-like unit is the short-time energy (STE) function. But the raw STE function is not reliable for segmentation due to significant local energy fluctuations [18]. We used group delay-based segmentation (GDS) to identify suprasegmental boundaries in the bird vocalisation. Group delay functions have been widely used for numerous speech and music process applications [1,20,20–22]. It is shown that if the signal is minimum phase, the group delay function resolves the peaks and valleys of the spectrum well [17]. If the short-term energy function is thought of as a magnitude spectrum, an equivalent minimum phase signal can be derived. The peaks and valleys of group delay function of this signal will now correspond to the peaks and valleys

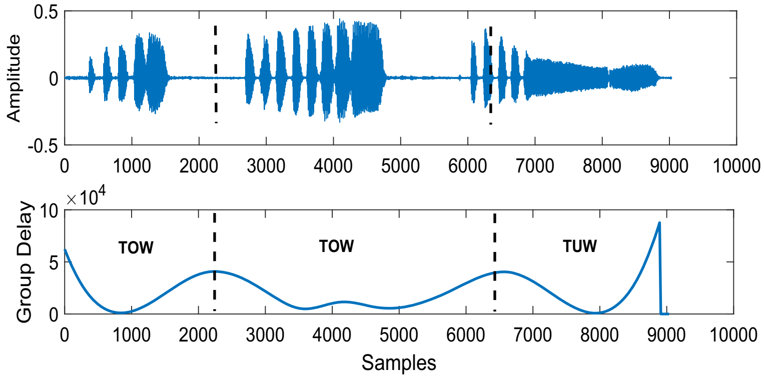

Fig. 3 Vocalization of red-wattled lapwing (upper pane). Group delay-based segmentation and corresponding labels are shown (Colour figure online)

in the short-term energy function. The location of the peaks in the minimum phase group delay function approximately corresponds to the syllable-like boundaries. In general, the number of syllables present is equal to the number of voiced segments. The algorithm for segmentation of continuous speech using this approach is explained in [12], which essentially smoothens the energy contour and removes the local energy fluctuation. For example, the segments detected using group delay-based algorithm for an audio file are shown in Fig. 2 with its computed group delay. Successive phones are combined based on the number of segments obtained from group delay-based segmentation, and lexicons are modified. By adopting the new boundary information, the lexicon [T-OW-T-OW-T-UW] in Table 1 is updated to [TOW-TOW-TUW]. Thus, lexicons of all the species are updated by combining phonemes, based on the outcome of group delay-based segmentation.

2.2 DNN-Based Acoustic Modelling

The standard Kaldi framework for DNN-based acoustic modelling [14] is implemented. The training of monophone model (GMM-HMM) and triphone model is experimented successively. In the next phase, triphone model is trained with LDA and MLLT followed by training on feature space maximum likelihood linear regression (fMLLR) adapted features. Finally, DNN-HMM hybrid model is trained. Maximum likelihood linear regression computes a set of transformations and reduces the mismatch between the initial model set and the adaptation data.

2.3 Deep Learning Methods on Acoustic Cues

The proposed scheme is compared with DNN, CNN and LSTM techniques. For DNN and LSTM frameworks, MFCCs are frame-wise computed from the audio file with a frame size of 10 ms and hop size of 5 ms. Our DNN architecture is based on three hidden layered feed-forward neural networks with 2048 nodes per layer. Relu has been

Table 2 CNN architecture for the experiment

Layer (type)	Output Shape
conv2d_1 (Conv2D)	(None, 150, 150, 32)
max_pooling2d_1 (Maxpooling)	(None, 75, 75, 32)
conv2d_2 (Conv2D)	(None, 75, 75, 64)
max_pooling2d_2 (Maxpooling)	(None, 37, 37, 64)
conv2d_3 (Conv2D)	(None, 37, 37, 32)
max_pooling2d_3 (Maxpooling)	(None, 18, 18, 32)
flatten_1 (Flatten)	(None, 10368)
dense_1 (Dense),	(None, 128)
dense_2 (Dense)	(None, 64)
dense_3 (Dense)	(None, 10)

chosen as the activation function for hidden layers and softmax function for the output layer. The training is performed in 100 epochs with a batch size of 10 using AdaMax optimization.

CNN architecture contains a deep architecture using repeated several convolution layers followed by max-pooling to process spectrograms. The detailed specifications of the CNN are given in Table 2. Spectrograms are generated from the vocalization using frame-size of 10 ms and hop size of 5 ms. The experiment is carried out with 100 epochs with a batch size of 32. The LSTM models are implemented with two stacked LSTM layers (888 nodes), followed by a dropout layer and the dense output layer. Tanh and sigmoid are used as the activation for cell state and output, respectively. Adam is chosen as the optimization algorithm for training. LSTM-RNNs can capture long-range temporal dependencies by overcoming the vanishing gradient problem in conventional approach [25].

3 Performance Evaluation

Test audio files are collected from widely used Xeno-canto bird sound database [29] of Xeno-canto foundation. [3] The test material consists of 563 audio recordings of 10 species. The files are refined such that one vocalization (approximately 1 s–3 s) is there in each audio file. The audio files are sampled at 16,000 Hz. A total of 325 files are used for testing and rest for training. The distribution on the number of audio recordings used in the experiment is listed in Table 1.

The primary and the most tiresome step in modelling the system was the data preparation stage [4,30]. Each bird species has several audio files, each file being several minutes long. These files are trimmed to 1 sec to 3 sec by ensuring one localization exists in the file. In the data preparation phase, transcriptions obtained from the CMUSphinx updated with GDS are forwarded as the lexicon in the DNN-based acoustic modelling. The experiment with GMM-HMM monophone and triphone models is performed in successive phases. During DNN training, training is started

[3] www.xeno-canto.org (Xeno-canto)

Table 3 Classification accuracy for the experiments

Sl.No	Method	Accr.(%)
1	Monophone (GMM-HMM)	61.00
2	Triphone (GMM-HMM, Delta -Delta)	66.50
3	Triphone (GMM-HMM, LDA+ MLLT)	68.50
4	Triphone (GMM-HMM, LDA+ MLLT+SAT)	70.50
5	Hybrid DNN-HMM + GDS	94.46
6	MFCC-DNN	72.92
7	Spectrogram-CNN	92.30
8	MFCC-LSTM	77.80

from the labelled frames, generated by a GMM-HMM system. In building the hybrid DNN-HMM system, the HMM states are considered to be the learning targets. The DNN-HMM model is trained using fMLLR-adapted features. The training of DNN-HMM is performed through the cross-entropy criterion in 100 epochs with the learning rate varying from 0.015 to 0.002.

The authors [16] emphasize the need for more training data in the visual representation-based approaches. It is stated that CNN needs a large size of data to achieve better results since it is not successful enough in low data sizes [13]. Since the initial results for CNN and LSTM frameworks were very low, with the training data specified in Table 1, additional training data of 595 (750 s) and 1235 (1560 s) for CNN and LSTM, respectively, were used for getting the results reported in Table 3.

4 Analysis of Results

The results of the entire experiment are given in Table 3. The accuracy reported for the conventional monophone model with GMM-HMM is 61% with 39-dim MFCC. The experiment is extended to the triphone model and triphone model with LDA, MLLT and SAT subsequently. As expected, triphone modelling showed an improvement of 5.5% in overall accuracy. The recognition accuracy obtained for GMM-HMM with triphone (LDA+MLLT) and triphone (LDA+MLLT+SAT) is 68.5% and 70.5%, respectively. It shows the significance of model adaptation in the proposed task. The architectural choice of DNN-HMM framework improved the system performance tremendously with an accuracy of 94.46%. Hybrid DNN-HMM model outperforms best-performing GMM-HMM model with an improvement of 23.96%. GMMs in HMM has a severe shortcoming that they are statistically inefficient for modelling data that lie on or near a nonlinear manifold in the data space [8]. Overall classification accuracy of 72.92%, 92.30%, 77.80% is obtained for DNN, CNN and LSTM, respectively. From Table III, it is observed that the hybrid DNN-HMM outperforms both CNN and LSTM frameworks

The confusion matrix for the MFCC-LSTM and hybrid DNN-HMM is given in Tables 5 and 6, respectively. For the first approach, class-wise accuracy of Eurasian

Table 4 Precision (P), recall (R) and F1 measure

SL.No	Class	DNN			CNN			LSTM			DNN-HMM		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
1	House crow (HC)	0.75	1.00	0.86	0.98	1.00	0.99	0.87	0.94	0.90	1.00	1.00	1.00
2	Mallard duck (MD)	1.00	0.30	0.46	1.00	0.97	0.98	0.84	0.87	0.85	1.00	0.90	0.95
3	Asian koel (AK)	0.59	0.59	0.59	1.00	0.89	0.94	0.68	0.70	0.69	0.91	0.86	0.88
4	Eurasian owl (EO)	0.85	0.58	0.69	0.98	1.00	0.99	1.00	0.38	0.54	1.00	0.90	0.95
5	House sparrow (HS)	0.91	0.91	0.91	1.00	0.94	0.97	0.89	0.89	0.89	1.00	0.97	0.98
6	Blue jay (BJ)	0.37	0.95	0.53	1.00	0.95	0.97	0.58	0.95	0.72	0.82	0.90	0.86
7	Red-wattled lapwing (RL)	1.00	0.85	0.91	0.87	1.00	0.93	1.00	0.30	0.46	0.86	0.95	0.90
8	Grey go-away (GG)	1.00	0.93	0.96	0.57	1.00	0.73	0.59	0.81	0.69	0.93	1.00	0.96
9	Indian peafowl (IP)	0.64	0.75	0.69	1.00	0.97	0.99	0.89	0.92	0.90	0.83	0.94	0.88
10	Western wood pewee (WW)	0.87	0.43	0.58	1.00	0.47	0.64	0.72	0.93	0.81	1.00	1.00	1.00
	Macro average	0.80	0.73	0.72	0.94	0.92	0.91	0.81	0.77	0.75	0.94	0.94	0.94

Table 5 Confusion matrix for MFCC-LSTM approach

	HC	MD	AK	EO	HS	BJ	RL	GG	IP	WW
HC	**47**	2	0	0	1	0	0	0	0	0
MD	0	**26**	0	0	3	1	0	0	0	0
AK	0	0	**26**	0	0	3	0	0	3	5
EO	0	0	11	**15**	0	0	0	14	0	0
HS	2	2	0	0	**31**	0	0	0	0	0
BJ	0	1	0	0	0	**19**	0	0	0	0
RL	1	0	1	0	0	7	**6**	0	1	4
GG	2	0	0	0	0	1	0	**22**	0	2
IP	2	0	0	0	0	0	0	1	**33**	0
WW	0	0	0	0	0	2	0	0	0	**28**

Diagonal values are bolded

Table 6 Confusion matrix for hybrid DNN-HMM approach (GDS)

	HC	MD	AK	EO	HS	BJ	RL	GG	IP	WW
HC	50	0	0	0	0	0	0	0	0	0
MD	0	**27**	0	0	0	0	0	2	1	0
AK	0	0	**32**	0	0	0	0	0	5	0
EO	0	0	2	**36**	0	0	2	0	0	0
HS	0	0	0	0	**34**	1	0	0	0	0
BJ	0	0	0	0	0	**18**	1	0	1	0
RL	0	0	0	0	0	1	**19**	0	0	0
GG	0	0	0	0	0	0	0	**27**	0	0
IP	0	0	0	0	0	2	0	0	**34**	0
WW	0	0	0	0	0	0	0	0	0	**30**

Diagonal values are bolded

owl and red-wattled Lapwing is less than 50%. But for the proposed approach, all the classes report accuracy higher than 80% with 100% accuracy for four species. Misclassification errors of red-wattled lapwing and Eurasian owl reduced significantly in the hybrid DNN-HMM model.

The precision, recall and F1 measure of the experiments are given in Table 4. Average precision, recall and F1 measure of the proposed approach are 0.94, 0.94 and 0.94, respectively. Average precision, recall and F1-measure for the four approaches are shown in Fig. 4 for a comparison. Average F1 measure for DNN-, CNN- and LSTM-based framework is 0.72, 0.91 and 0.75, respectively.

It can be seen that the proposed approach shows a neck and neck performance with CNN-based approach with slight mileage for the proposed scheme. It performs significantly better than DNN and LSTM models. It is worth pointing out that the better performance of the proposed approach is achieved with reasonably less amount of training data as compared to other deep learning models. Group delay segmentation-based suprasegmental lexicon modelling is also explored efficiently in the task. The

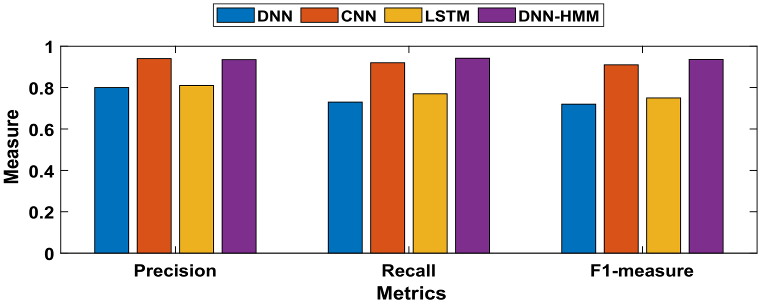

Fig. 4 The performance metrics for four phases

results show that human speech recognition can potentially be applied for "transcribing" bird vocalizations and thereby classification.

5 Conclusion and Future Scope

The potential of the standard speech recognition framework is experimented in transcribing bird vocalizations and thereby classifying them to appropriate species. We have addressed the bird call classification through transcription approach using DNN-based acoustic models. Hybrid DNN-HMM has the advantage of adopting strong learning power from DNNs and the sequential modelling ability from HMMs. The performance is evaluated using a dataset with 10 species. The recognition accuracy of the approach is 94.46%. Results show that the performance of the proposed method outperforms spectrogram-CNN, MFCC-LSTM and MFCC-DNN frameworks. As a future work, we would like to extend the proposed framework to analyse vocalization with more species. Besides, the performance can be studied in adverse conditions, where sounds due to rain and wind cause disturbances in real conditions. In some instances, birds make vocalisations simultaneously, which makes automated sound analysis even more difficult. Instead of isolated conditions, multiple bird vocalization can also be studied as a future scope of the proposed work. In [31], the authors summarize recent progresses made in deep learning-based acoustic models. The paper shows the success of the models such as RNNs and CNNs that can effectively exploit variable-length contextual information and their various combinations with other models. The proposed framework can potentially be benefited from those insights.

Declaration

Data Availability The datasets generated during and/or analysed during the current study are available in the Xeno-canto bird sound repository-www.xeno-canto.org (Xeno-canto)

References

1. R. Ajayakumar, R. Rajan , Predominant instrument recognition in polyphonic music using gmm-dnn framework. pp. 1–5 (2020). 10.1109/SPCOM50965.2020.9179626

2. J.Bonada, R. Lachlan, M. Blaauw, Bird song synthesis based on hidden markov models. In *Proc. of International Conference on Spoken Language Processing* pp. 2582–2586 (2016)
3. D. Chakraborty, P. Mukker, P. Rajan, A.D. Dileep, Bird call identification using dynamic kernel based support vector machines and deep neural networks. In *Proc. of 15th IEEE International Conference on Machine Learning and Applications (ICMLA)* pp. 280–285 (2016)
4. W. Chu, D.T. Blumstein, Noise robust bird song detection using syllable pattern-based hidden Markov models. In *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing* pp. 345–348 (2011)
5. D.B. Efremova, M. Sankupellay, D.A. Konovalov, Data-efficient classification of birdcall through convolutional neural networks transfer learning. Digital image computing: techniques and applications pp. 1–8 (2019)
6. D. Gelling, Bird song recognition using GMMs and HMMs. Masters Project Dissertation, Department of Computer Science, University of Sheffield (2001)
7. A.Harma, Automatic identification of bird species based on sinusoidal modeling of syllables. In *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing* **5**, V–545 (2003)
8. G. Hinton, L. Deng, D. Yu, G.E. Dahl, A.r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath, et al., Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. Signal Processing Magazine pp. 82–97 (2012)
9. P. Jancovic, M. Kokue, M. Zakeri, M. Russell, Bird species recognition using HMM-based unsupervised modelling of individual syllabls with incoparated duration modelling. In *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing* pp. 559–563 (2016)
10. P. Jancovic, M. Köküer, Bird species recognition using unsupervised modeling of individual vocalization elements. IEEE/ACM Transactions on Audio, Speech, and Language Processing pp. 932–947 (2019). 10.1109/TASLP.2019.2904790
11. P. Jancovic, M. Köküer, M.Russell, Bird species recognition from field recordings using HMM-based modelling of frequency tracks pp. 8252–8256 (2014)
12. T.N.K. Prasad, H.A. Murthy, Automatic segmentation of continuous speech using minimum phase group delay functions. Speech Commun. **42**, 429–446 (2004)
13. M. Kaya, S.H. Bilge, Deep metric learning: a survey. Symmetry **11**(9), 1–26 (2019)
14. I. Kipyatkova, A. Karpov, DNN-based acoustic modeling for Russian speech recognition using Kaldi. In *Proc. of the International Conference on Speech and Computer, LNAI 9811* pp. 1–8 (2016)
15. E. Knight, K. Hannah, G. Foley, C. Scott, R. Brigham, E. Bayne, Recommendations for acoustic recognizer performance assessment with application to five common automated signal recognition programs. Avian Conservation and Ecology 12(2) (2017)
16. C. Liua, L. Fengb, G. Liuc, H. Wangd, S. Liub, Bottom-up broadcast neural network for music genre classification. Pattern Recognit. Lett. pp. 1–7 (2019)
17. H. Murthy, B. Yegnanarayana, Group delay functions and its application to speech processing. Sadhana **36**(5), 745–782 (2011)
18. T. Nagarajan, H. Murthy, M. Rajesh, Group delay based segmentation of spontaneous speech into syllable-like units. EURASIP J. Appl. Signal Processi. pp. 2641–2625 (2004)
19. I. Potamitis, S. Ntalampiras, K.R. Olaf Jahn, Automatic bird sound detection in long real-field recordings: applications and tools. Appl. Acoust. pp. 1–9 (2014)
20. R. Rajan, H.A. Murthy, *Group Delay Based Melody Monopitch Extraction from Music* (In: Proceedings of the IEEE Int. Conf. on Audio, Speech and Signal Processing Pp, 2013), pp. 186–190
21. R. Rajan, H. AMurthy, Music genre classification by fusion of modified group delay and melodic features. In Proc. of National Conference on Communications (2017)
22. R. Rajan, H.A. Murthy, Two-pitch tracking in co-channel speech using modified group delay functions. Speech Commun. **89**, 37–46 (2017)
23. P. Somervuo, A. Harma, Analyzing bird song syllables on the self-organizing map. In *Proc. of the Workshop on Self-Organizing Maps, Hibikino, Japan* (2003)
24. D. Stowell, M. Wood, Y. Stylianou, H. Glotin, Bird detection in audio: A survey and a challenge. In Proc. of IEEE International Workshop on Machine Learning for Signal Processing Salerno, Italy pp. 1–6 (2016)
25. C.P. Tang, K. Chui, Y. u, Z. Zeng, K. Wong, Music genre classification using a hierarchical long short term memory (LSTM) model. In Proc. of International Conference on Information Retrieval,Yokohama,Japan pp. 521–526 (2018)

26. A. Thakur, V. Abrol, P. Sharma, P. Rajan, Compressed convex spectral embedding for bird species classification. In Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing pp. 261–265 (2018)
27. A. Thakur, V. Abrol, P. Sharma, P. Rajan, Deep convex representations: Feature representations for bioacoustics classification. In Proc. of International Conference on Spoken Language Processing pp. 2127–2131 (2018)
28. A. Thakur, V. Abrol, P. Sharma, P. Rajan, Local compressed convex spectral embedding for bird species identification. The J. Acoust. Soc. Am. **143**, 3819–3828 (2018)
29. W.P. Vellinga, R. Planqué, Working notes of conference and labs of the evaluation forum (CELF) (2015)
30. S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland, The HTK book (2002)
31. D. Yu, J. Li, Recent progresses in deep learning based acoustic models. IEEE/CAA J. Automat. Sinica **4**(3), 396–409 (2017). https://doi.org/10.1109/JAS.2017.7510508

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.