Separating Overlapping Bat Calls with a Bi-directional Long Short-term Memory Network

Kangkang Zhang
Northeast Normal University
Tong Liu
Northeast Normal University
Shengjing Song
Northeast Normal University
Xin Zhao
Northeast Normal University
Shijun Sun
Northeast Normal University
Walter Metzner
University of California Los Angeles
Jiang Feng
Jilin Agricultural University
Ying Liu (liuy252@nenu.edu.cn)
Northeast Normal University https://orcid.org/0000-0003-3359-0716

Methodology

Keywords: Bat vocalizations, bioacoustics, BLSTM, call mixtures, deep neural networks, overlapping calls, sound separation

Posted Date: March 30th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-19737/v1

Version of Record: A version of this preprint was published at Integrative Zoology on May 30th, 2021. See the published version at https://doi.org/10.1111/1749-4877.12549.

Abstract

Background: Acquiring clear and usable audio recordings is critical for acoustic analysis of animal vocalizations. Bioacoustics studies commonly face the problem of overlapping signals, which can impede the structural identification of vocal units, but the issue is often ignored, as there is currently no satisfactory solution. This study presents a bi-directional long short-term memory (BLSTM) network to separate overlapping bat calls and reconstruct waveforms. The separation quality was evaluated using seven temporal-spectrum parameters. The applicability of this method for bat calls was assessed using six different species. In addition, clustering analysis was conducted with separated echolocation calls from each bat species.

Results: All syllables in the overlapping calls were separated with high robustness across species. A comparison between the seven temporal-spectrum parameters showed no significant difference and negligible deviation between the extracted and original calls, indicating high separation quality. Clustering analysis also produced an accuracy of 93.8%, suggesting the reconstructed waveforms could be reliably used.

Conclusions: The study extends the application of deep neural network to separate overlapping animal sound. These results suggest the proposed technique is a convenient and automated approach for separating overlapping calls using a BLSTM network. This powerful deep neural network approach has the potential to solve complex problems in bioacoustics.

Background

The structural identification of vocal units is essential in animal acoustic studies for sound feature analysis, sound emitter recognition, and species identification and monitoring. However, animal monitoring, both in the field and in the laboratory, often involves problems caused by the overlapping of different vocal units in time and frequency space, which prevents the acoustic components from being suitable for parameter analysis. As a result, the separation of overlapping sounds is an important task in bioacoustic signal processing. However, existing analysis software often struggles to process overlapping calls and previous research on the acoustic identification of animals primarily focuses on extracting target signals from background noise for species classification or population monitoring [1–4]. The process of separating overlapping calls from mixed sounds has received little attention to date and researchers conventionally abandon sounds that overlap in both time and frequency, requiring an extension of the experimental period to obtain sufficient non-overlapping recordings [5]. As such, an effective method for successfully and automatically separating overlapping calls would be of significant interest and benefit to animal researchers.

Previous studies using deep neural networks have produced promising results for automated sound recognition in complex acoustic environments for animal species recognition and classification [5–7]. However, in this study, we consider the more difficult task of separating different types of syllables from

overlapping calls and reconstructing sound waves from these separated signals. Existing techniques used for animal sound separation often require prohibitive quantities of labelled data. For example, multiple-instance machine learning (MIML) algorithms were proposed for use in sound feature extraction and species identification in birds [1]. However, this technique requires a cropped mask of a signal segment (without overlap) in order to extract each syllable.

Deep learning networks have been applied to bioacoustic studies but have primarily been used for classification. For instance, convolutional bidirectional recurrent neural networks (CBRNNs) have been used to identify the presence of bird calls in audio samples [4]. Acoustic features were learned by the network (a classifier) and the presence or absence of a bird call was output as an indicator. Similarly, convolutional neural networks (CNNs) have been used to predict the presence of a search-phase bat echolocation call in spectrograms [2]. To our knowledge, the use of deep learning techniques to separate animal calls that overlap in both time and frequency space has yet to be reported.

Multiple studies have been conducted using deep learning-based supervised speech separation with humans. Early systems included shallow models that performed a linear transformation of given mixture features during the prediction time interval. This has included Gaussian mixture models [8], support vector machines [9], and non-negative matrix factorization [10]. However, in real-world scenarios, the mapping relationship between mixture signals and sources is typically a nonlinear transformation. Nonlinear models, such as deep neural networks (DNNs), are therefore highly applicable because of their ability to identify nonlinear structures in audio signals [11–13]. Additionally, recurrent neural networks (RNNs) that exhibit the temporal behaviour of a time sequence can be trained to predict time-frequency masks for target signals and separate sources from a mixed waveform [14]. Specifically, long short-term memory (LSTM) networks, a variation of RNN models that exhibit strong learning capabilities and simple construction, have been widely used for word and continuous speech recognition [15–17]. By concatenating two separate LSTM networks, bidirectional LSTMs (BLSTMs) can predict each element of a sequence based on past and future context and can naturally account for the temporal dynamics of speech. These models are typically faster and more accurate than standard RNNs in frame-by-frame phoneme classification [18]. In addition, the BLSTM network can compensate for exploding and vanishing gradient issues that can occur during the training of standard RNN models [19]. At present, BLSTMs have achieved state-of-the-art performance for speech recognition [13, 20], natural language processing [21, 22], and speaker-independent speech separation [23]. As such, a BLSTM model was selected in this study for overlapping bat call separation.

Echolocating bats have two vocal repertoires, echolocation calls for orientation and a variety of communication calls for social activities [24–26]. Recordings from both field and laboratory studies indicate that utterances from individual bats often overlap in both time and frequency, which provides an excellent template for research on overlapping sound separation in animals. The primary objective of this study is to develop a technique for separating two target signals (echolocation and socialization calls) from mixtures of acoustic sounds. Although deep learning has been employed in the acoustic

classification of multiple species, including nonhuman primates [27], birds [4], whales [28], and bats [2, 3], the deep neural network in the present study is applied for single channel source separation.

Both overlapping and non-overlapping calls (of both echolocation and communication types) were recorded from each of the collected bat species studied in our previous work. We developed a BLSTM network and used the recorded non-overlapping calls to train the model. Recorded overlapping calls were input to the trained model and separated. Independent waveforms were then reconstructed for each separated signal. The separated signals were validated by comparing the temporal-spectrum parameters between separated calls and the originally recorded (non-overlapping) calls from each species. Finally, clustering analysis was conducted to classify the separated echolocation calls based on bat species, which provided a practical application of the proposed technique.

Materials And Methods

Sound recording and data preparation

Species selection and sound sources. Echolocation calls from bats are primarily composed of constant frequency (CF) components and frequency modulated (FM) components. Social calls are composed of CF, FM, and noise-burst (NB) components. FM calls have short pulse durations and wide bandwidths. As such, they overlap with social calls less in time but more in frequency. In contrast, CF calls have long pulse durations and narrow bandwidths. They overlap with social calls more in time but less in frequency. In consideration of the varied overlapping patterns found in bat calls, we selected both CF bats (*Rhinolophus ferrumequinum*, *Hipposideros armiger*, and *Rhinolophus pusillus*) and FM bats (*Vespertilio sinensis*, *Myotis macrodactylus*, and *Ia io*) to test the separation capabilities of the proposed network, including six different species to test method generalizability.

Source sound files from *V. sinensis*, *M. macrodactyllus*, *R. ferrumequinum*, *R. pusillus*, and *H. armiger* were collected from previous studies in our lab (Appendix S1). Sound files for *Ia io* were selected from unpublished data as follows. Bats captured from the field were housed in a husbandry room with abundant food and fresh water. During each sound recording experiment, 4–5 bats were transferred to a temporary cage. Sound recordings were collected using the Avisoft UltraSoundGate 116H (Avisoft Bioacoustics, Berlin, Germany) and a condenser ultrasound microphone (CM16/CMPA, Avisoft Bioacoustics). The sampling frequency was set to 375 kHz at 16 bits. The recording experiment lasted five days in order to acquire a sufficient number of recordings, beginning at 18:00 and finishing at 6:00 the following morning. Appendix S1 shows sample numbers and locations for the bats, as well as the total duration of sound files selected for the study.

Sound analysis. The total duration of recorded sound files (i.e., original recording files) used for each bat species were shown in Appendix S1. We employed Avisoft-SASLab Pro (Version 5.2.12, Avisoft Bioacoustics, Berlin, Germany) to manually identify non-overlapping and overlapping syllables in echolocation and communication calls. These syllables and calls were described and classified following the nomenclature developed by Kanwal, Matsumura [29] and Ma, Kobayasi [30]. The recorded non-

overlapping calls were used for preparing training files of each call type and the recorded overlapping calls were used for separation.

Data preparation. Supervised machine learning algorithms use training samples to "learn" how to complete a task. The training phase in this study involved preparing clear and non-overlapping echolocation and communication calls, selected from original recording sounds. In this process, the BLSTM network learned features found in both call types.

Training samples consisted of randomly selected non-overlapping syllables in echolocation and communication calls from each bat species (in the original recordings), with signal-to-noise ratios (SNRs) above −20 dB. The echolocation training files contained 1,300–6,240 pulses and the communication training files contained 780–1,800 syllables (Appendix S1). Although the quantity of selected syllables varied between studies, the data was sufficient for model training. Efforts were made to include roughly equivalent quantities of each syllable type. Time intervals between syllables in the training files were consistent with those of the original recordings. The lengths of training files for echolocation and communication calls were the same for each bat species (Appendix S1).

Model training and call separation

Model structure and training stage. We developed a network with four BLSTM layers, followed by one feedforward layer (Fig. 1). Each BLSTM layer included one forward and one backward basic LSTM layer, both of which were added with dropout functions (tensorflow.nn.rnn_cell.DropoutWrapper). Each BLSTM layer contained 300 hidden cells and the feedforward layer corresponded to the embedding dimension (i.e., a 3D matrix with depth N=40 in this experiment). Stochastic gradient descent with a momentum of 0.9 and a fixed learning rate of 10^{-3} was used for training. The tanh activation function and the Adam optimizer were adopted to support adaptive learning rates and faster convergence. The structure and hyper-parameters for the model were designed based on the work of Hershey, Chen [20].

Fig. 1 The BLSTM model architecture and workflow graph.

The model was trained using the files for one bat species in each trial. Echolocation and communication call training files were loaded using the librosa (version 0.6.2) Python package. Frames from the two sound files were read and added together to create sound mixtures. Sound features used for training (log spectral magnitudes) were extracted from this mixture. The extraction process was completed using a short-time Fourier transform (STFT) with a Hamming window (length of 512 and shift of 256).

The mixture from each bat species was then segmented into 100-frame samples, all of which were divided into a training set and a validation set using a ratio of 2:1 (see Appendix S1 for detailed sample quantities). The training set, validation set, and indicator labels were combined and input to the model. The validation set was used to optimize tuning parameters and evaluate call separation performance. Indicator labels were set to 0 or 1, representing the two types of calls in the mixture. Ideal binary masks

were used to train the network and gradients were calculated using shuffled mini-batches (batch size of 128) from larger segments.

The output of this model was a set of embeddings that included learned features for both echolocation and communication calls. In this framework, the deep network assigned embedding vectors to each time-frequency bin in the spectrogram. The network then minimized the distance between embeddings dominated by the same call type in each bin while maximizing the distance between embeddings dominated by different call types. The output was then compared with the validation set and indicator labels to calculate loss, which was back propagated from the output to the input through each layer. Model weights and parameters were then updated based on the calculated loss and training was completed after sufficient iteration epochs.

Separation stage. In this stage, overlapping echolocation and communication calls were randomly selected from the original recordings to create a sound file of test sets, used for separation. The log spectral magnitudes of the overlapping calls were then extracted, combined into samples, and input to the trained model. The phases of calls extracted from the sound files were also saved for use in sound reconstruction. The trained model then output embeddings for each segment (100 frames) in a process similar to the training stage. Embeddings were clustered using the k-means method from Scikit-learn (Version 0.20.0) to produce time-frequency masks. The number of clusters corresponded to the number of call types in the mixture (2 - echolocation and communication). These masks and the clustering method were then used to determine which parts of each segment in the overlapped calls would be preserved or neglected based on their correspondence to each call type. For example, if the maximum magnitudes were more likely to belong to echolocation calls, the related mask values were set to 1 and the others were set to 0, allowing the echolocation calls to be separated correctly. Finally, output calls were reconstructed using the inverse fast Fourier transform (IFFT) function numpy.fft.ifft in NumPy (Version 1.15.1). The IFFT transformed the magnitude into a wave using phase information saved at the beginning of the separation stage. The model produced two waveform files, each containing one call type. Additional detail concerning the sound separation algorithms can be found in the work of Hershey, Chen [20].

Model evaluation

The quality of reconstructed echolocation and communication calls was assessed by comparing their temporal-spectrum parameters to the non-overlapping calls selected from the original recording files (excluding training data). Avisoft-SASLab Pro was used for automatic parameter measurements of duration, bandwidth, peak frequency, minimum frequency, maximum frequency, starting frequency, and ending frequency. A t-SNE (t-distributed stochastic neighbour embedding - R3.6.1 package) analysis was adopted for dimensionality reduction. Two dimensions were extracted from these seven parameters for original and separated syllables and compared with one-way ANOVA (aov in R3.6.1) or two-sided Wilcoxon signed-rank tests (wilcox.test in R3.6.1), depending on their fit to a normal Gaussian distribution. The significance level was set to 0.05 for all tests. We adopted the root mean square error

(RMSE) to measure and avoid obscuring individual variations between reconstructed and original calls. Clustering analysis was conducted using the reconstructed echolocation calls from the six bat species, to assess whether the separated calls could be further used in species classification.

Results

The proposed algorithm performed well and achieved high accuracy in separating overlapping calls for each of the six species. The BLSTM model was iteratively trained until the training and validation losses reached a minimum. Loss is a summation of errors made with each sample in the training or validation sets and measures how well the model adapts during optimization. The validation loss function tended toward an asymptotic value, indicating the network had converged (Appendix S2 Fig. S1).

All echolocation and communication calls in the overlapping signals were correctly extracted during the separation procedure, regardless of their pulse duration or energy characteristics (see Table 1 and Fig. 2). In addition, low-intensity FM components in echolocation pulses were successfully extracted from three CF bat species (Figs 2D, 2E, and 2F).

Table 1. Separation results.

Species	Call type	Number of syllable types	Number of syllables in overlapping calls	Number of overlapping syllables	Number of separated syllables
Rhinolophus ferrumequinum	Echolocation	1	14	14	14
	Communication	4	8	8	8
Vespertilio sinensis	Echolocation	1	21	13	13
	Communication	4	8	8	8
Hipposideros armiger	Echolocation	1	28	19	19
	Communication	6	13	13	13
Myotis macrodactylus	Echolocation	1	54	36	36
	Communication	6	15	15	15
Rhinolophus pusillus	Echolocation	1	42	30	30
	Communication	6	10	10	10
Ia io	Echolocation	1	26	16	16
	Communication	4	11	11	11

Fig. 2 Spectrograms from original recordings of overlapping calls and calls separated by the BLSTM

network. The first graph represents each line of the original overlapping calls and the second and third graphs show the separated echolocation and communication calls, respectively.

A comparison of seven temporal-spectrum parameters from the separated calls and the original recorded non-overlapping calls showed no significant differences (Fig. 3 and Appendix S2 Table S1). In addition, parameter deviations in separated calls and original non-overlapping calls showed minimal root mean square error (RMSE) values for both echolocation and communication signals (Appendix S2 Fig. S2 and S3). Clustering analysis performed with separated echolocation calls produced an accuracy of 93.8% across species (Appendix S2 Fig. S4).

Fig. 3 Comparisons between the separated and original calls. Two principle components extracted from seven temporal-spectrum parameters were plotted, and statistical analysis were elaborated in Material and Methods. Results for echolocation and communication calls are shown in (A-F) and (G-L), respectively.

Discussion

The BLSTM network used in the present study achieved high accuracy in separating overlapping echolocation and communication calls from bats. The training and validation loss for the model also exhibited fast convergence and high robustness for bat vocalizations. In particular, the separated calls extracted by the proposed algorithm were reconstructed as waveforms with nearly the same quality as the non-overlapping calls, suggesting BLSTM networks to be useful tools for separating signals in future bioacoustic research, such as sound analysis, acoustic identification, species classification, and wild animal monitoring.

It was difficult to compare the performance of this algorithm with that of previous studies, primarily because of differences in the experimental procedure and no standard animal sound datasets as benchmarks. However, a comparison of temporal-spectrum parameters between separated calls and non-overlapping calls was included as an evaluation metric. The seven parameters used in this study are commonly used in bat studies to describe the temporal-spectrum features of syllables [25, 31]. Statistical results for this comparison showed no significant differences and small deviations in parameters between separated calls and original recordings, indicating the system was able to separate calls without affecting syllable quality. In addition, clustering analysis conducted with reconstructed echolocation calls was highly accurate (93.8%) for species classification, indicating that calls separated from overlapping signals could be used for further evaluation.

The BLSTM network exhibited good performance across all six bat species using both narrow and broad time-frequency calls. No species-specific a priori knowledge or particular acoustic sensor was directly encoded into the system, making it generalizable to other animal populations with additional training data. Although the dichotomy between communication and echolocation calls is relatively drastic, the proposed separation system has potential applications for even more complex tasks. One might wonder if the network could separate the overlaps of echolocation calls or social calls. If labelled calls under

different situations are available and enough for training data, this is the standard supervised separation problem. As mixtures of echolocation pulses and social call syllables are very common in bats, in the present study we selected such mixtures and our main purpose is to introduce a method that could separate syllables from mixtures, which has received little study so far.

While deep learning models generally perform better when provided with more data, training with bat calls requires fewer samples than human speech separation, in which available training sets can exceed hundreds of hours [12]. One possible reason for this may be the high signal-to-noise ratio (SNR) of bat sounds recorded with high-quality ultrasound devices. Previous studies have indicated that a high SNR can improve separation accuracy [32] and our results suggest this model was suitable for use with small, high-quality datasets. Although the sound data in this study were sampled in controlled lab conditions, producing recordings that were essentially free of background noise, acoustic analysis software might potentially optimize the separation further by excluding any background noise to increase the SNR.

Future studies will also assess the performance of this network for other animal species. Stereotypical patterns and clearly classifiable syllables have been observed in the vocalizations of birds, non-human primates, whales, dolphins, and several other species [33–35]. Features used in the proposed BLSTM were log spectral magnitudes, which can be acquired from any vocal sound. This could potentially lead to robust software that is not specific to a certain species or task, though limitations may exist. In addition to the quality and quantity of training samples, hyper-parameters (such as number of network layers or nodes in each layer) should be tuned in accordance with the data and tasks [36, 37].

Conclusion

A sound separation model was proposed for extracting bat calls, achieving excellent results. This is the first experimental evidence that the BLSTM model is suitable for separating overlapping bioacoustic signals. These results provided a new source for sound data analysis in animal acoustics research, which may contribute to sample sizes and improve efficiency. This study also demonstrates the potential of deep neural networks for applications to support the automatic monitoring of wild bat populations in the field, which is important for species preservation.

Declarations

Ethics approval and consent to participate

All experimental procedures complied with the ABS/ASAB guidelines for the Use of Animals in Research and were approved by the Committee on the Use and Care of Animals at the Northeast Normal University (approval number: NENU-W-2014–101).

Acknowledgements

We are grateful to Dr. Yanhong Xiao of the Experimental Center of the School of Environment at Northeast Normal University, for her assistance in acquiring the experimental materials. We thank LetPub (www.letpub.com) for its linguistic assistance during the preparation of this manuscript.

Funding

This work was supported by the National Natural Science Foundation of China, Grant Nos. 31770429 (to YL) and 31670390 (to JF), the Natural Science Foundation of Jilin, Grant No. 20180101263JC (to YL), the Program for Introducing Talents to Universities, Grant No. B16011 (to YL), and the National Program for “1000 Talent Plan for High-Level Foreign Experts”, Grant No. WQ20142200259 (to WM).

Author contributions

KKZ, WM, YL, and JF developed the methodology; KKZ, TL, SJS, and XZ collected the data and trained the BLSTM network; KKZ and SJS analysed the results and composed the manuscript.

Competing interests

The authors declare that no competing interests exist.

Consent for publication

The authors have given final approval for publication.

Availability of data and materials

All data and code have been deposited in GitHub: https://github.com/zkkandrew/batcallseparation

Authors’ information

[1]Jilin Provincial Key Laboratory of Animal Resource Conservation and Utilization, Northeast Normal University, Changchun, China. [2]School of Environment, Northeast Normal University, Changchun, China. [3]Department of Integrative Biology and Physiology, University of California, Los Angeles, California, USA. [4]Collage of Animal Science and Technology, Jilin Agricultural University, Changchun, China.

References

1. Briggs F, Lakshminarayanan B, Neal L, Fern XZ, Raich R, Hadley SJK, et al. Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach. 2012;131(6):4640-50.10.1121/1.4707424
2. Aodha OM, Gibb R, Barlow KE, Browning E, Firman M, Freeman R, et al. Bat detective—Deep learning tools for bat acoustic signal detection. PLOS Computational Biology. 2018;14(3):156869

3. Walters CL, Freeman R, Collen A, Dietz C, Brock Fenton M, Jones G, et al. A continental-scale tool for acoustic identification of European bats. Journal of Applied Ecology. 2012;49(5):1064-74.https://doi.org/10.1111/j.1365-2664.2012.02182.x
4. Adavanne S, Drossos K, Cakir E, Virtanen T. Stacked convolutional and recurrent neural networks for bird audio detection. european signal processing conference. 2017:1729-33.10.23919/EUSIPCO.2017.8081505
5. Priyadarshani N, Marsland S, Castro I. Automated birdsong recognition in complex acoustic environments: a review. Journal of Avian Biology. 2018;49(5):jav-01447.https://doi.org/10.1111/jav.01447
6. Redgwell RD, Szewczak JM, Jones G, Parsons S. Classification of echolocation calls from 14 species of bat by support vector machines and ensembles of neural networks. Algorithms. 2009;2(3):907-24
7. Sprengel E, Jaggi M, Kilcher Y, Hofmann T, editors. Audio based bird species identification using deep learning techniques. LifeCLEF 2016; 2016
8. Kim G, Lu Y, Hu Y, Loizou PC. An algorithm that improves speech intelligibility in noise for normal-hearing listeners. The Journal of the Acoustical Society of America. 2009;126(3):1486-94.10.1121/1.3184603
9. Wang Y, Wang D. Towards Scaling Up Classification-Based Speech Separation. IEEE Transactions on Audio, Speech, and Language Processing. 2013;21(7):1381-90.10.1109/TASL.2013.2250961
10. Grais EM, Erdogan H, editors. Spectro-temporal post-smoothing in NMF based single-channel source separation. Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European; 2012: IEEE;
11. Nugraha AA, Liutkus A, Vincent E. Deep Neural Network Based Multichannel Audio Source Separation. Signals Commun Techn. 2018:157-85.10.1007/978-3-319-73031-8_7
12. Wang D, Chen J. Supervised Speech Separation Based on Deep Learning: An Overview. IEEE Transactions on Audio, Speech, and Language Processing. 2018;26(10):1702-26.10.1109/Taslp.2018.2842159
13. Marchi E, Ferroni G, Eyben F, Gabrielli L, Squartini S, Schuller B, editors. Multi-resolution linear prediction based features for audio onset detection with bidirectional LSTM neural networks. 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP); 2014: IEEE;
14. Weninger F, Hershey JR, Le Roux J, Schuller B, editors. Discriminatively trained recurrent neural networks for single-channel speech separation. Proceedings 2nd IEEE Global Conference on Signal and Information Processing, GlobalSIP, Machine Learning Applications in Speech Processing Symposium, Atlanta, GA, USA; 2014
15. Eck D, Graves A, Schmidhuber J. A new approach to continuous speech recognition using LSTM recurrent neural networks. Technical Report. 2003
16. Beringer N, editor Human language acquisition methods in a machine learning task. Eighth International Conference on Spoken Language Processing; 2004

17. Graves A, Beringer N, Schmidhuber J, editors. A Comparison Between Spiking and Differentiable Recurrent Neural Networks on Spoken Digit Recognition. international conference on modelling identification and control; 2004
18. Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks. 2005;18(5):602-10.https://doi.org/10.1016/j.neunet.2005.06.042
19. Makino S. Audio Source Separation: Springer; 2018.
20. Hershey JR, Chen Z, Roux JL, Watanabe S, editors. Deep clustering: Discriminative embeddings for segmentation and separation. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2016 20-25 March 2016.10.1109/ICASSP.2016.7471631
21. Wöllmer M, Eyben F, Graves A, Schuller B, Rigoll GJCC. Bidirectional LSTM networks for context-sensitive keyword detection in a cognitive virtual agent framework. 2010;2(3):180-90
22. Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:150801991. 2015
23. Li C, Zhu L, Xu S, Gao P, Xu B, editors. CBLDNN-Based Speaker-Independent Speech Separation Via Generative Adversarial Training. international conference on acoustics, speech, and signal processing; 2018
24. Kunz TH, Fenton MB. Bat ecology: University of Chicago Press; 2005.
25. Gillam E, Fenton MB. Roles of acoustic social communication in the lives of bats. Bat Bioacoustics: Springer; 2016. p. 117-39.https://doi.org/10.1007/978-1-4939-3527-7_5
26. Luo B, Huang X, Li Y, Lu G, Zhao J, Zhang K, et al. Social call divergence in bats: a comparative analysis. Behavioral Ecology. 2017;28(2):533-40.10.1093/beheco/arw184
27. Pozzi L, Gamba M, Giacoma C. The Use of Artificial Neural Networks to Classify Primate Vocalizations: A Pilot Study on Black Lemurs. Am J Primatol. 2010;72(4):337-48.10.1002/ajp.20786
28. Shamir L, Yerby C, Simpson R, Benda-Beckmann AMv, Tyack P, Samarra F, et al. Classification of large acoustic datasets using machine learning and crowdsourcing: Application to whale calls. The Journal of the Acoustical Society of America. 2014;135(2):953-62.https://doi.org/10.1121/1.4861348
29. Kanwal JS, Matsumura S, Ohlemiller K, Suga N. Analysis of acoustic elements and syntax in communication sounds emitted by mustached bats. The Journal of the Acoustical Society of America. 1994;96(3):1229-54.https://doi.org/10.1121/1.410273
30. Ma J, Kobayasi K, Zhang S, Metzner W. Vocal communication in adult greater horseshoe bats, Rhinolophus ferrumequinum. Journal of comparative physiology A, Neuroethology, sensory, neural, and behavioral physiology. 2006;192(5):535-50.https://doi.org/10.1007/s00359-006-0094-9
31. Jin L, Wang J, Zhang Z, Sun K, Kanwal JS, Feng J. Postnatal development of morphological and vocal features in Asian particolored bat, Vespertilio sinensis. Mammalian Biology. 2012;77(5):339-44.10.1016/j.mambio.2012.05.001

32. Weng C, Yu D, Seltzer ML, Droppo J. Deep neural networks for single-channel multi-talker speech recognition. IEEE Transactions on Audio, Speech, and Language Processing. 2015;23(10):1670-9.10.1109/Taslp.2015.2444659
33. Naguib M, Riebel K. Bird song: a key model in animal communication. Encyclopedia for language and linguistics. 2006;2:40-53
34. Filatova OA, Deecke VB, Ford JKB, Matkin CO, Barrett-Lennard LG, Guzeev MA, et al. Call diversity in the North Pacific killer whale populations: implications for dialect evolution and population history. Animal behaviour. 2012;83(3):595-603.https://doi.org/10.1016/j.anbehav.2011.12.013
35. Herzing DL. Clicks, whistles and pulses: Passive and active signal use in dolphin communication. Acta Astronautica. 2014;105(2):534-7.10.1016/j.actaastro.2014.07.003
36. Lecun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436.10.1038/nature14539
37. Goodfellow I, Bengio Y, Courville A. Deep Learning: The MIT Press; 2016.

Figures

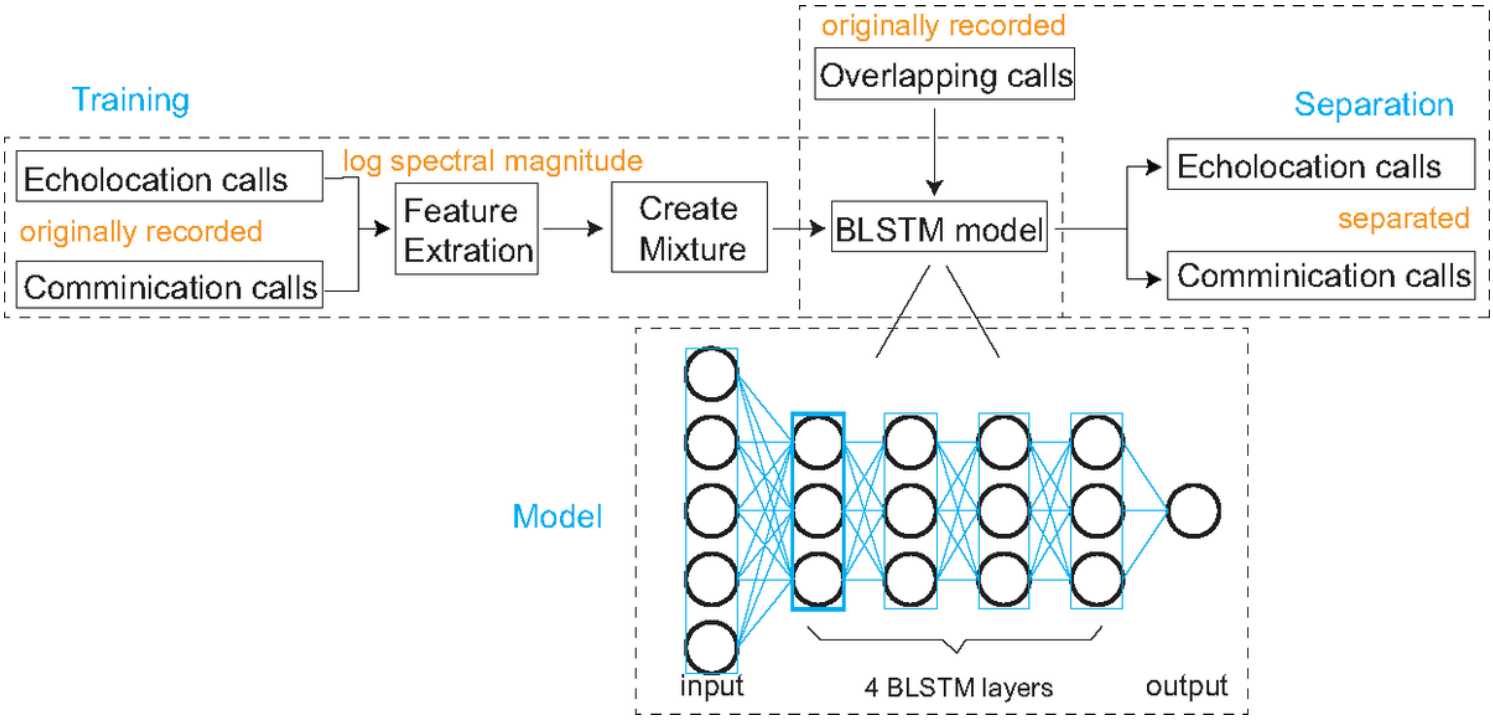

Figure 1

The BLSTM model architecture and workflow graph.

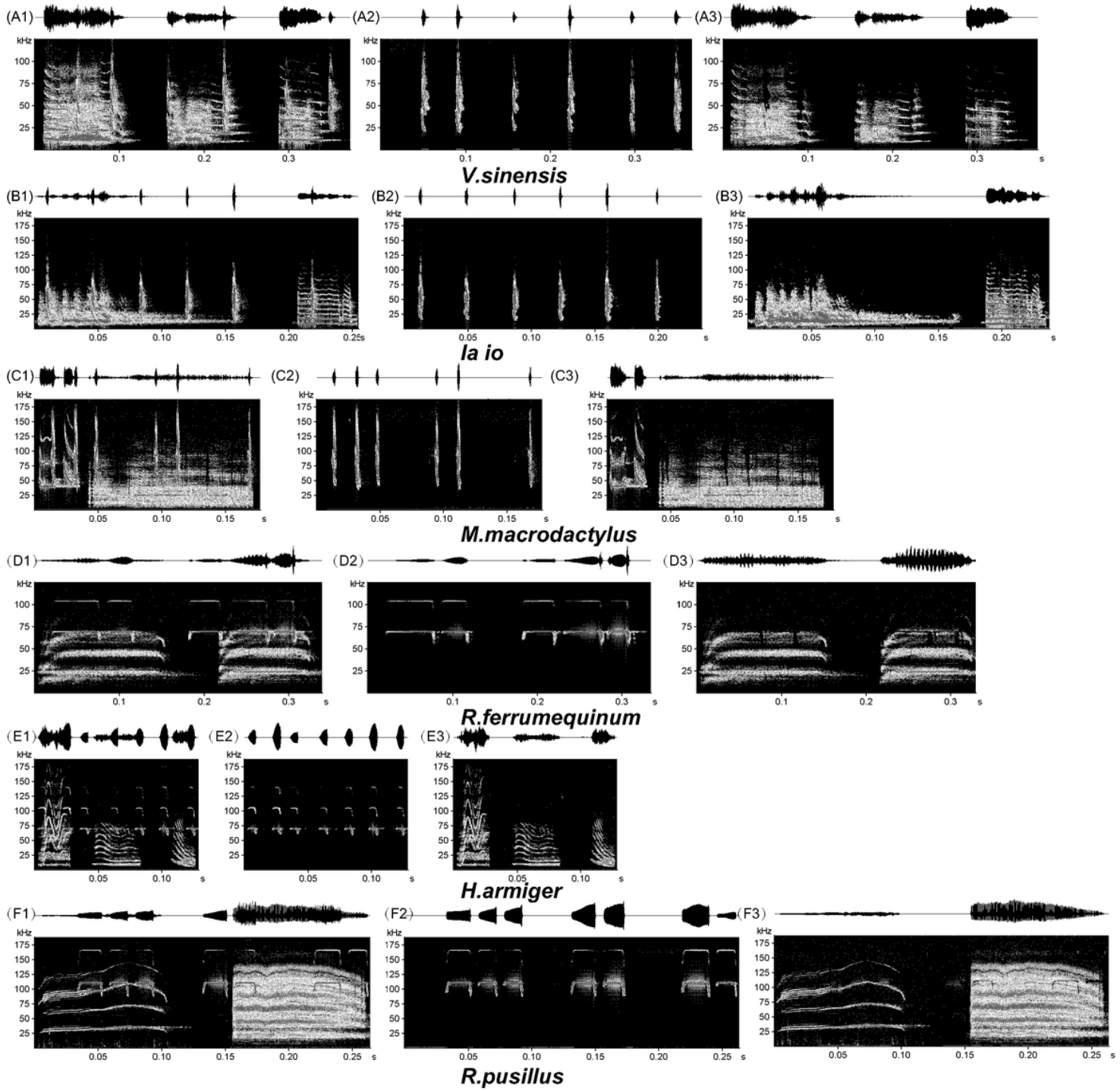

Figure 2

Spectrograms from original recordings of overlapping calls and calls separated by the BLSTM network. The first graph represents each line of the original overlapping calls and the second and third graphs show the separated echolocation and communication calls, respectively.

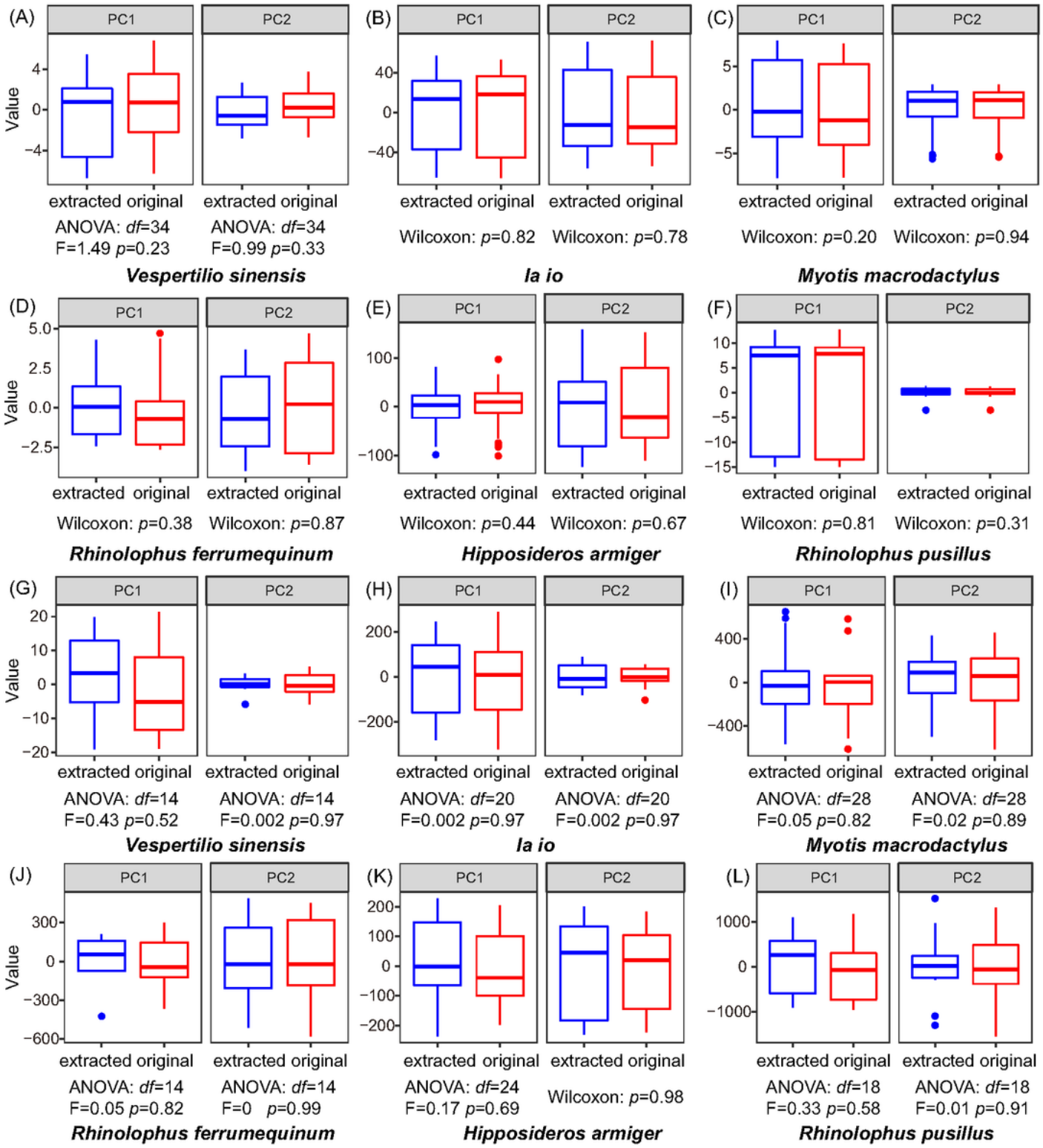

Figure 3

Comparisons between the separated and original calls. Two principle components extracted from seven temporal-spectrum parameters were plotted, and statistical analysis were elaborated in Material and Methods. Results for echolocation and communication calls are shown in (A-F) and (G-L), respectively.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AppendixS1summaryofbasicdata.xlsx
- AppendixS2FigS3.eps
- AppendixS2FigS1.eps
- AppendixS2FigS4.eps
- AppendixS2FigS2.eps
- AppendixS2statisticalresults.docx