Separating overlapping bat calls with a bi-directional long short-term memory network

Using deep neural network to separate overlapping bat calls

Kangkang Zhang[1], Tong Liu[1], Shengjing Song[1], Xin Zhao[1], Shijun Sun[2], Walter Metzner[3], Jiang Feng[1,4*], Ying Liu[1*]

[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.

[*]Correspondence

E-mail: liuy252@nenu.edu.cn (YL)

E-mail: fengj@nenu.edu.cn (JF)

Abstract

Acquiring clear and usable audio recordings is critical for acoustic analysis of animal vocalizations. Bioacoustics studies commonly face the problem of overlapping signals, 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 waveform audio sounds. 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 population. Results showed that 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 of the separated echolocation calls also produced an accuracy of 93.8%, suggesting the reconstructed waveform sounds could be reliably used. 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.

Author summary

In recent years, the development of recording techniques and devices in animal acoustic experiment and population monitoring has led to a sharp increase in the volume of sound data. However, the collected sound would be overlapped because of the existence of multiple individuals, which laid restrictions on taking full advantage of experiment data. Besides, more convenient and automatic methods are needed to cope with the large datasets in animal acoustics. The echolocation calls and communication calls of bats are variable and often overlapped with each other both in the recordings from field and laboratory, which provides an excellent template for research on animal sound separation. Here, we firstly solved the problem of overlapping calls in bats successfully based on deep neural network. We built a network to separate the overlapping calls of six bat species. All the syllables in overlapping calls were separated and we found no significant difference between the separated syllables with non-overlapping syllables. We also demonstrated an instance of applying our method on species classification. Our study provides a useful and efficient model for sound data processing in acoustic research and the proposed method has the potential to be generalized to other animal species.

Introduction

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, wild 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 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, 6]. 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 [6-8]. 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. Convolutional neural networks (CNNs) have been used to predict the presence of a search-phase bat echolocation call in spectrograms. This binary classification problem was used to detect the presence of bats [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 [9], support vector machines [10], and non-negative matrix factorization [11]. 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 [12-14]. Additionally, recurrent neural networks (RNNs) that exhibit the temporal behavior of a time sequence can be trained to predict time-frequency masks for target signals and separate sources from a mixed waveform [15]. 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 [16-18]. 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 [19]. In addition, the BLSTM network can compensate for exploding and vanishing gradient issues that can occur during the training of standard RNN models [20]. At present, BLSTMs have achieved state-of-the-art performance for speech recognition [14, 21], natural language processing [22, 23], and speaker-independent speech separation [24]. As such, a

BLSTM model was selected in this study for overlapping bat call separation.

Echolocating bats have two vocal repertoires, stereotypical echolocation calls for orientation and a variety of communication calls for social activities [25-27]. 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 leaning has been employed in the acoustic classification of multiple species, including nonhuman primates [28], birds [4], whales [5], and bats [2, 3], the goal of the present study is distinct from these previous cases in which deep neural networks were primarily used as classifiers.

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 sound files were then reconstructed for each separated signal. The correctness of these separated signals was measured by comparing the temporal-spectrum parameters between separated calls and the initially recorded (non-overlapping) calls from each species. Finally, clustering analysis was conducted to classify the bats using separated echolocation calls, which provided a practical application of the proposed technique.

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. Training loss for this model decreased significantly in the first epoch. The validation loss function tended toward an asymptotic value, indicating the training algorithm had converged (S2 Fig). The BLSTM model converged slightly faster when training with CF bat samples (as opposed to FM samples).

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 1). In addition, low-intensity FM components in echolocation pulses were successfully extracted from three CF bat species (Figs 1d, 1e, and 1f).

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	Echolocation	1	54	36	36

macrodactylus	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 1. 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 2 and S3 Table). In addition, parameter deviations in separated calls and original non-overlapping calls showed minimal RMSE values for both echolocation and communication signals (Fig 3 and Fig 4). Clustering analysis performed with separated echolocation calls produced an accuracy of 93.8% across species (Fig 5).

Fig 2. Comparisons between the separated and original calls. Two principle components extracted from seven temporal-spectral parameters were used in the study. Results for echolocation and communication calls are shown in (A-F) and (G-L), respectively.

Fig 3. A comparison of deviations for separated and original echolocation calls. The RMSE value is shown under each plot. The vertical axis represents values for each parameter and the horizontal axis represents the number of syllables measured. The red triangles represent separated calls and the blue dots represent original calls. Abbreviations include duration (duration), Fstart (starting frequency), Fend (ending frequency), Fpeak (peak frequency), Fmin (minimum frequency), Fmax (maximum frequency), and bandw (bandwidth).

Fig 4. A comparison of deviations in separated and original communication calls. The RMSE value is shown under each plot. The vertical axes and abbreviations are the same in Fig 3.

Fig 5. Clustering analysis for six bat species based on their separated echolocation calls. Overlapping echolocation signals cannot be used for species identification until after separation.

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 waveform files 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. 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-spectral features of syllables [26, 29]. 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 to synthesize initial data.

The BLSTM network exhibited good performance across all six bat species using both narrow and broad time-frequency calls. It also successfully separated different syllable types from both overlapping echolocation and communication calls (Table 1, Fig 1). 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 other species, as such mixtures are very common in bats. In the future, more complex emitter-independent separation could be conducted using the proposed system, such as combinations of echolocation or social calls from other animals.

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 [13]. 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 [30] 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 could potentially optimize the separation further by excluding any background noise that was present in the signal.

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 [31-33]. 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. The model could also be generalized to other animals, though limitations may exist. In addition to the quality and quantity of training samples, hyper-parameters must be tuned in accordance with the data [34, 35].

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 animal vocalization research, including species classification and speech separation.

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 (S1 Table). 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. S1 Table shows sample numbers and locations for the bats, as well as the total duration of sound files selected for the study. 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-2010–101).

Sound analysis. The total duration of recorded sound files (i.e., original recording files) used for each bat species is shown in S1 Table. We employed Avisoft-SASLab Pro (Version 5.2.12, Avisoft Bioacoustics, Berlin, Germany) to 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 (36) and Ma, Kobayasi (37). 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” the steps required for completing 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 (S1 Table). 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 (S1 Table).

Model training and call separation

Model structure and training stage. We developed a network with four BLSTM layers, followed by one feedforward layer (Fig 6). 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 (21).

Fig 6. 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 S1 Table 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 (2016).

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 neighbor 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.

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.

Author contributions

Conceptualization: Kangkang Zhang, Walter Metzner.

Data curation: Tong Liu, Shengjing Song, Xin Zhao.

Formal Analysis: Kangkang Zhang, Tong Liu.

Methodology: Kangkang Zhang, Ying Liu, Jiang Feng.

Software: Kangkang Zhang, Shijun Sun.

Funding acquisition: Ying Liu, Jiang Feng, Walter Metzner.

Supervision: Jiang Feng.

Visualization: Kangkang Zhang, Tong Liu.

Writing – original draft: Kangkang Zhang, Ying Liu.

Writing – review & editing: Ying Liu, Walter Metzner.

Competing interests

The authors have declared that no competing interests exist.

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. doi: 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. doi: 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. doi: 10.23919/EUSIPCO.2017.8081505.
5. 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. doi: https://doi.org/10.1121/1.4861348.
6. Priyadarshani N, Marsland S, Castro I. Automated birdsong recognition in complex acoustic environments: a review. Journal of Avian Biology. 2018;49(5):jav-01447. doi: https://doi.org/10.1111/jav.01447.
7. 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.
8. Sprengel E, Jaggi M, Kilcher Y, Hofmann T, editors. Audio based bird species identification using deep learning techniques. LifeCLEF 2016; 2016.
9. 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. doi: 10.1121/1.3184603.
10. Wang Y, Wang D. Towards Scaling Up Classification-Based Speech Separation. IEEE Transactions on Audio, Speech, and Language Processing. 2013;21(7):1381-90. doi: 10.1109/TASL.2013.2250961.
11. 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.
12. Nugraha AA, Liutkus A, Vincent E. Deep Neural Network Based Multichannel Audio Source Separation. Signals Commun Techn. 2018:157-85. doi: 10.1007/978-3-319-73031-8_7. PubMed PMID: WOS:000441018800008.
13. 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. doi: 10.1109/Taslp.2018.2842159. PubMed PMID: WOS:000438718300001.
14. 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.
15. 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.

16. Eck D, Graves A, Schmidhuber J. A new approach to continuous speech recognition using LSTM recurrent neural networks. Technical Report. 2003.

17. Beringer N, editor Human language acquisition methods in a machine learning task. Eighth International Conference on Spoken Language Processing; 2004.

18. 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.

19. Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks. 2005;18(5):602-10. doi: https://doi.org/10.1016/j.neunet.2005.06.042.

20. Makino S. Audio Source Separation: Springer; 2018.

21. 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.

22. 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.

23. Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:150801991. 2015.

24. 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.

25. Kunz TH, Fenton MB. Bat ecology: University of Chicago Press; 2005.

26. Gillam E, Fenton MB. Roles of acoustic social communication in the lives of bats. Bat Bioacoustics: Springer; 2016. p. 117-39.

27. 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. doi: 10.1093/beheco/arw184.

28. 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. doi: 10.1002/ajp.20786. PubMed PMID: WOS:000276124300007.

29. 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. doi: 10.1016/j.mambio.2012.05.001. PubMed PMID: WOS:000309023900005.

30. 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. doi: 10.1109/Taslp.2015.2444659. PubMed PMID: WOS:000356518100012.

31. Naguib M, Riebel K. Bird song: a key model in animal communication. Encyclopedia for language and linguistics. 2006;2:40-53.

32. 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. doi: https://doi.org/10.1016/j.anbehav.2011.12.013.

33. Herzing DL. Clicks, whistles and pulses: Passive and active signal use in dolphin communication. Acta Astronautica. 2014;105(2):534-7. doi: 10.1016/j.actaastro.2014.07.003.

34. Lecun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436. doi: 10.1038/nature14539. PubMed PMID: 26017442.

35. Goodfellow I, Bengio Y, Courville A. Deep Learning: The MIT Press; 2016.

36. 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. doi: https://doi.org/10.1121/1.410273.

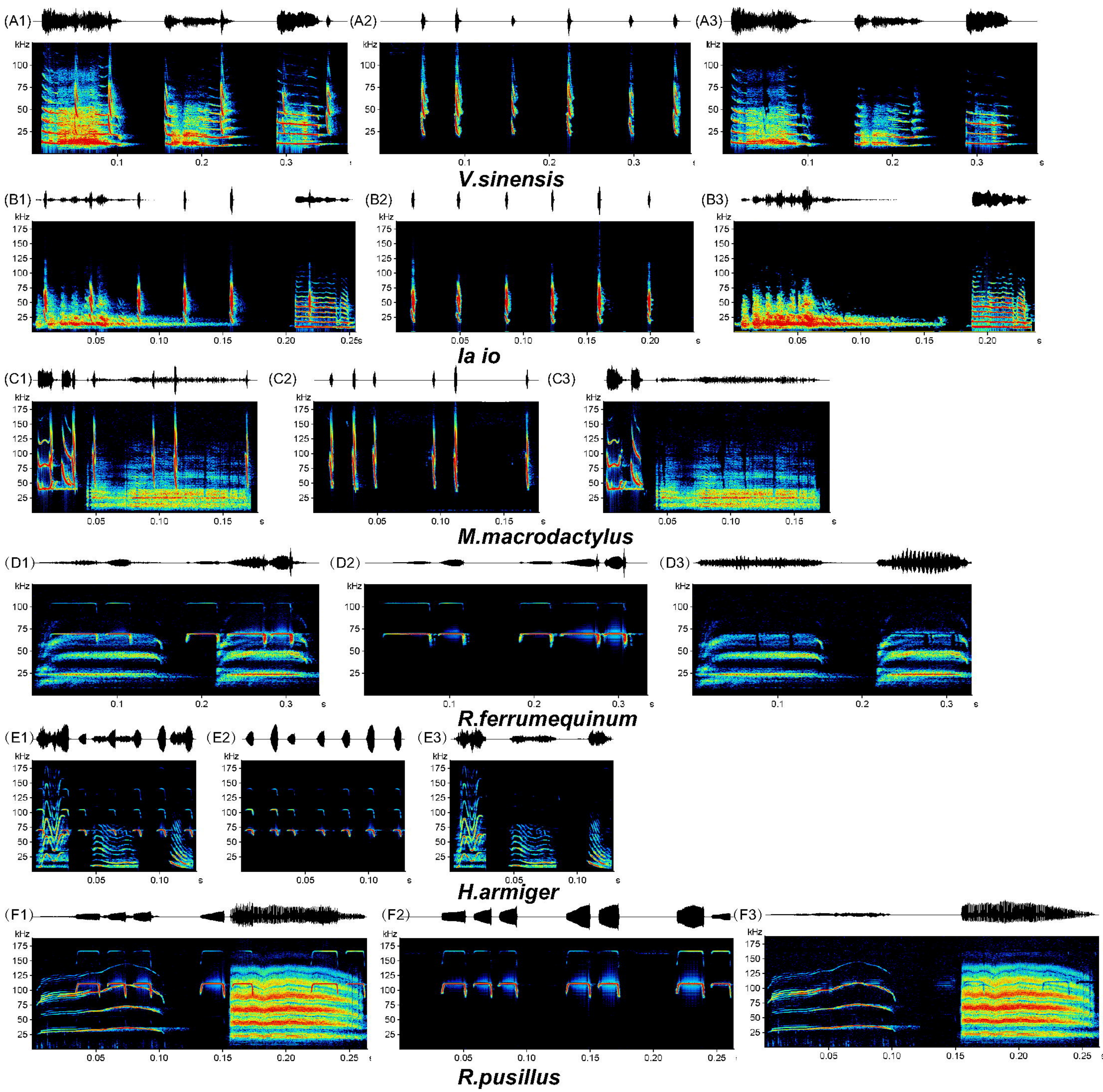
37. Ma J, Kobayasi K, Zhang S, Metzner W. Vocal communication in adult greater horseshoe bats, Rhinolophus ferrumequinum. Journal of comparative physiology A, Neuroethology, sensory, neural, and behavioral physiology. 2006;192(5):535-50. doi: https://doi.org/10.1007/s00359-006-0094-9. PubMed PMID: 16418857.

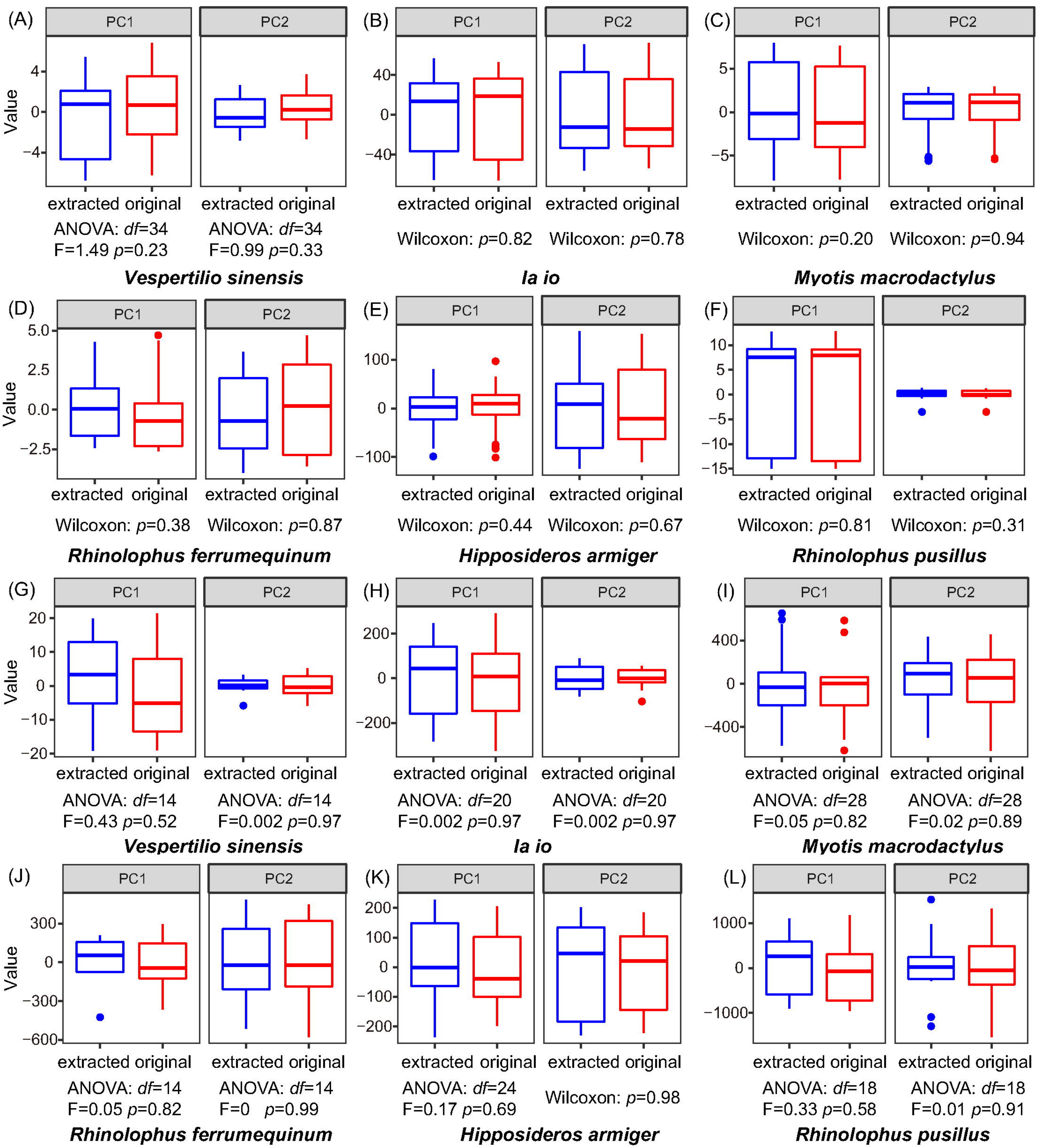
Supporting information

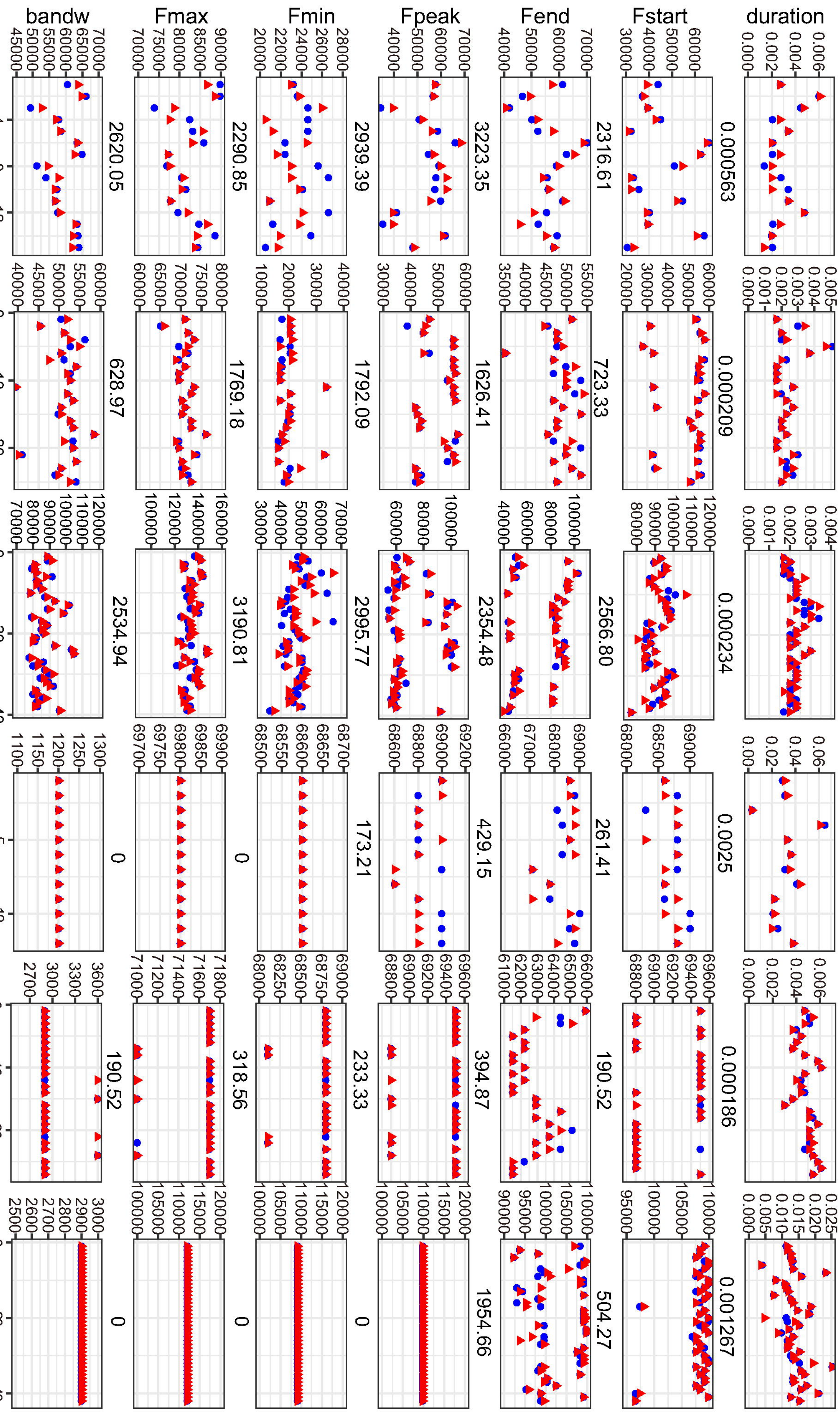
S1 Table. A summary of calls used for model training.

S2 Fig. Training loss and validation loss during model training.

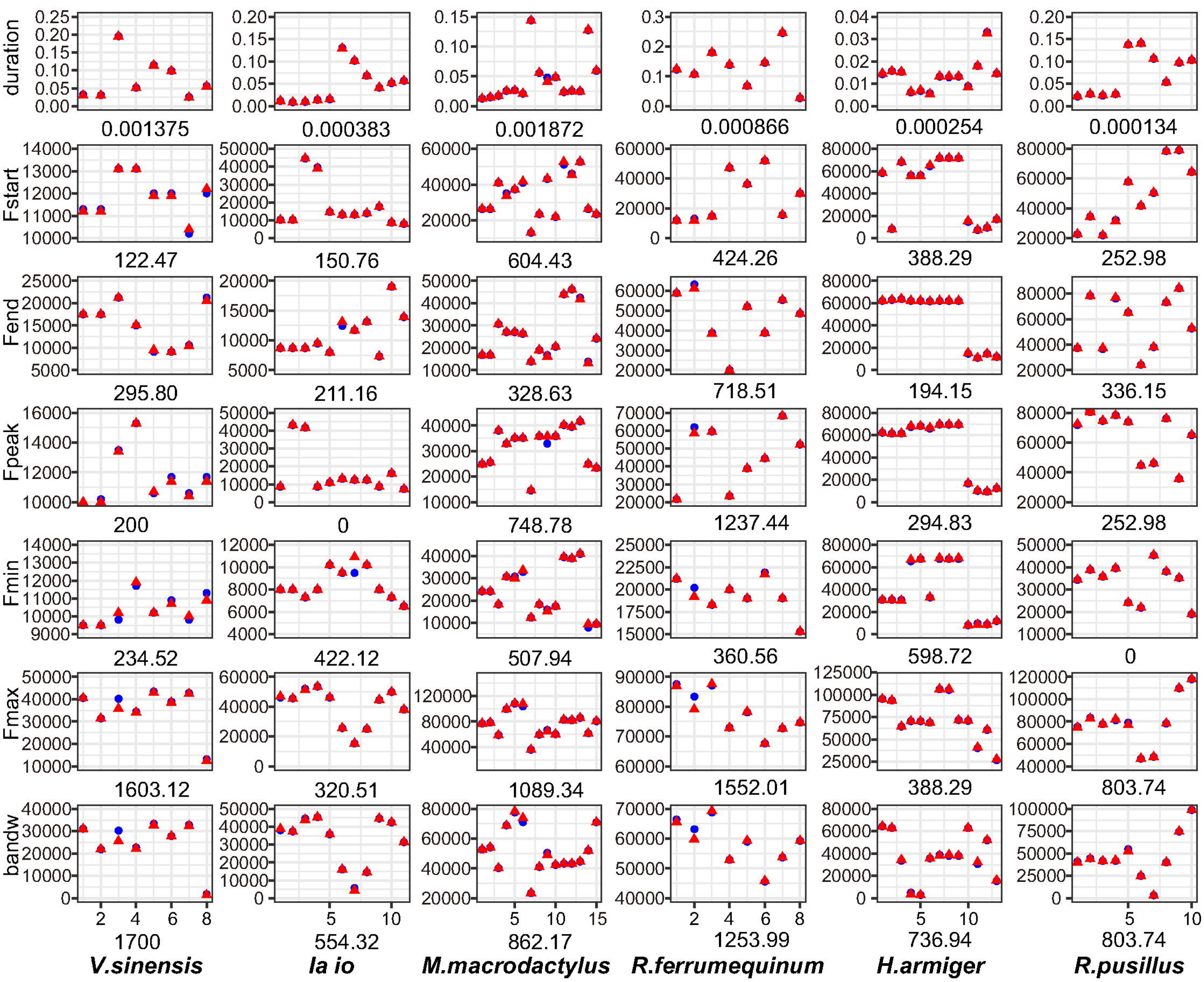
S3 Table. Statistical comparisons of principle components extracted from seven parameters. No significant differences were observed between parameters for separated and original syllables. A one-way ANOVA was used to test the normal distributed data and a two-sided Wilcoxon signed-rank test was used to assess the data that did not conform well to a normal distribution.

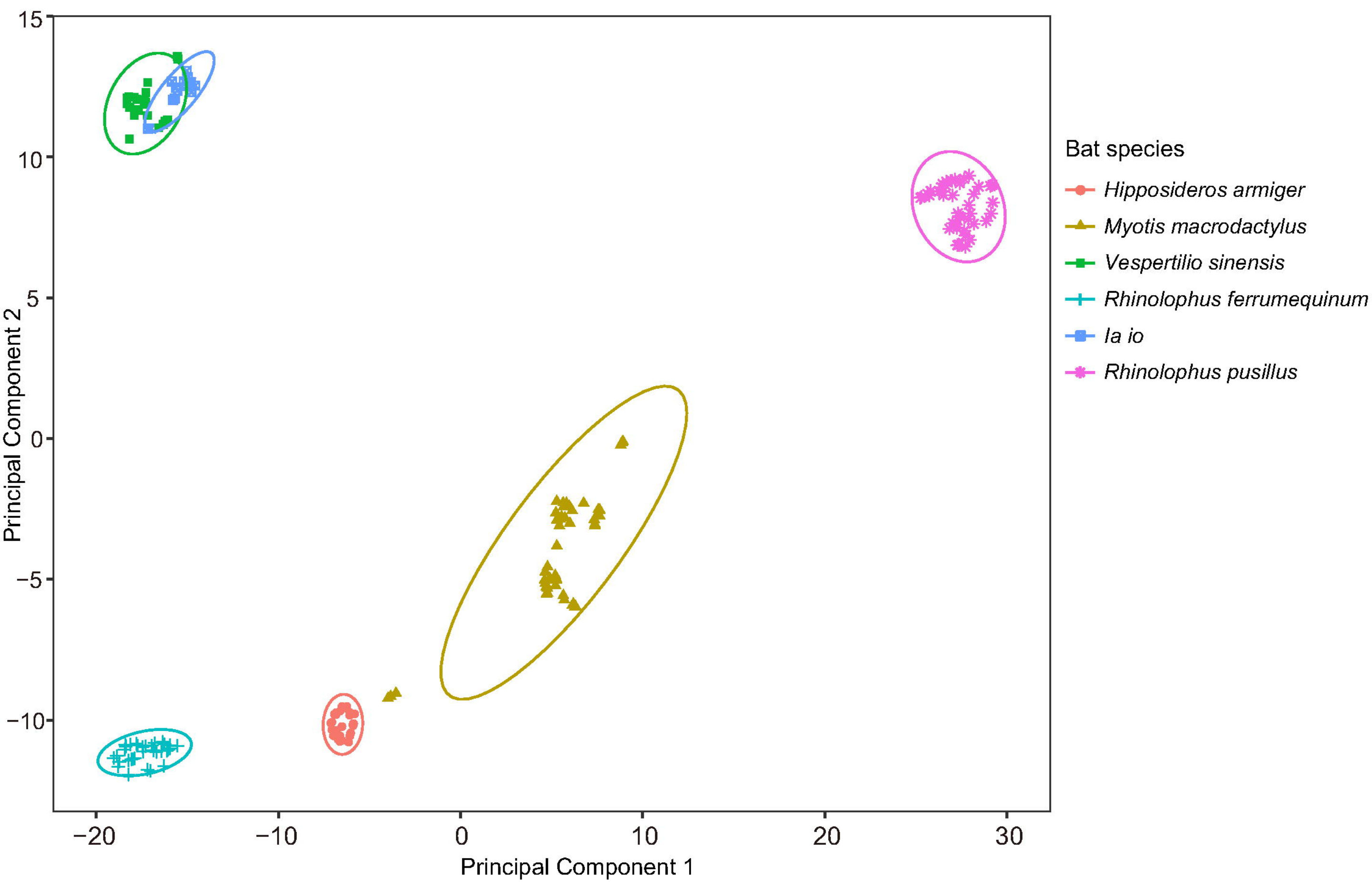

(A1)
(A2)
(A3)
V.sinensis
(B1)
(B2)
(B3)
Ia io
(C1)
(C2)
(C3)
M.macrodactylus
(D1)
(D2)
(D3)
R.ferrumequinum
(E1)
(E2)
(E3)
H.armiger
(F1)
(F2)
(F3)
R.pusillus
kHz
s

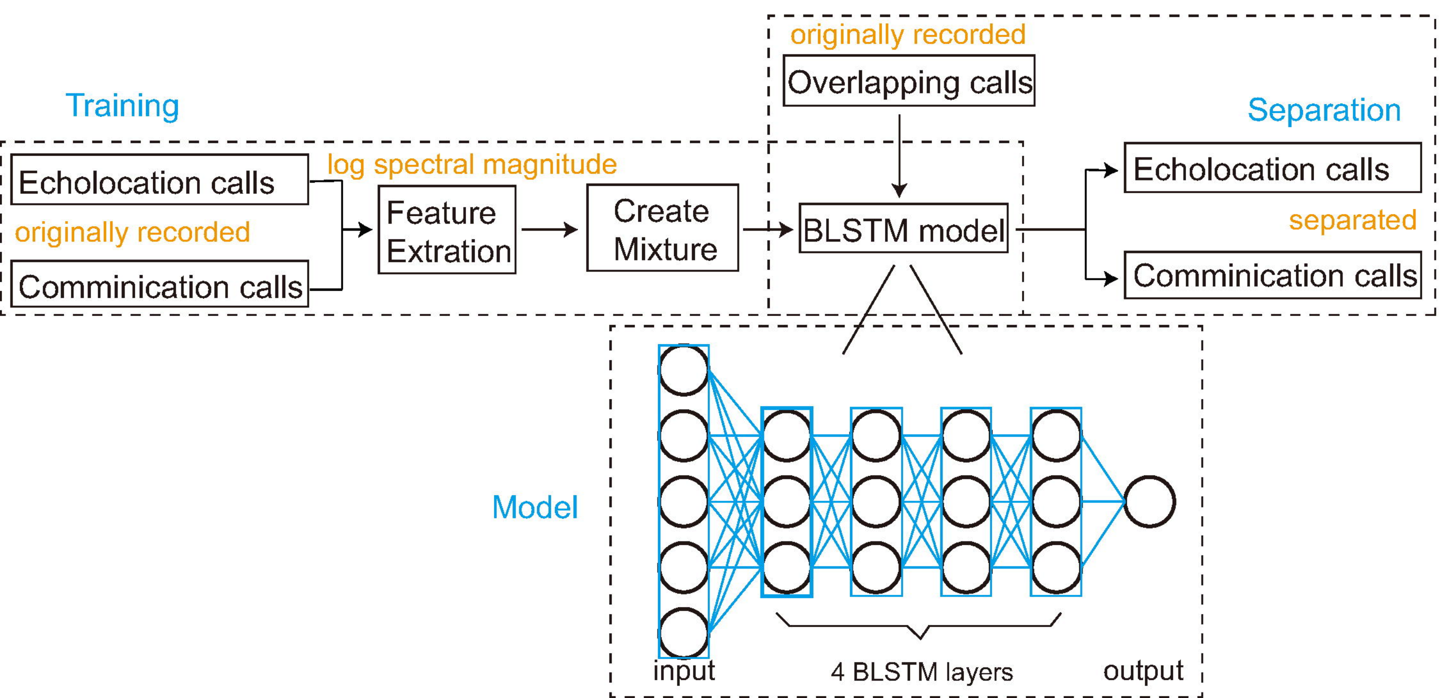

(A)
PC1
PC2
Value
extracted original
ANOVA: df=34
F=1.49 p=0.23
ANOVA: df=34
F=0.99 p=0.33
Vespertilio sinensis
(B)
Wilcoxon: p=0.82
Wilcoxon: p=0.78
Ia io
(C)
Wilcoxon: p=0.20
Wilcoxon: p=0.94
Myotis macrodactylus
(D)
Wilcoxon: p=0.38
Wilcoxon: p=0.87
Rhinolophus ferrumequinum
(E)
Wilcoxon: p=0.44
Wilcoxon: p=0.67
Hipposideros armiger
(F)
Wilcoxon: p=0.81
Wilcoxon: p=0.31
Rhinolophus pusillus
(G)
ANOVA: df=14
F=0.43 p=0.52
ANOVA: df=14
F=0.002 p=0.97
Vespertilio sinensis
(H)
ANOVA: df=20
F=0.002 p=0.97
ANOVA: df=20
F=0.002 p=0.97
Ia io
(I)
ANOVA: df=28
F=0.05 p=0.82
ANOVA: df=28
F=0.02 p=0.89
Myotis macrodactylus
(J)
ANOVA: df=14
F=0.05 p=0.82
ANOVA: df=14
F=0 p=0.99
Rhinolophus ferrumequinum
(K)
ANOVA: df=24
F=0.17 p=0.69
Wilcoxon: p=0.98
Hipposideros armiger
(L)
ANOVA: df=18
F=0.33 p=0.58
ANOVA: df=18
F=0.01 p=0.91
Rhinolophus pusillus



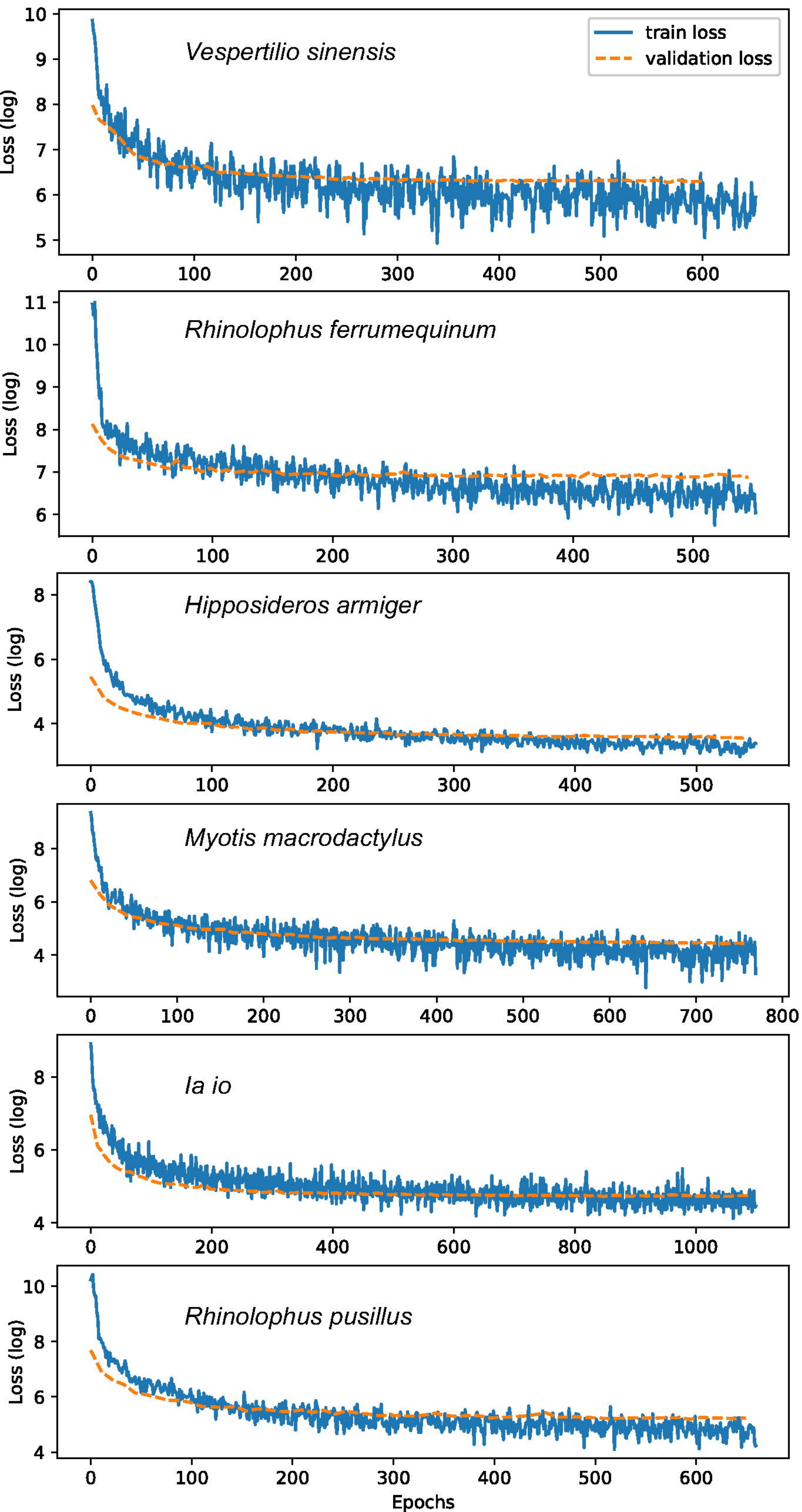
duration
Fstart
Fend
Fpeak
Fmin
Fmax
bandw
V.sinensis
Ia io
M.macrodactylus
R.ferrumequinum
H.armiger
R.pusillus
0.000563
2316.61
3223.35
2939.39
2290.85
2620.05
2037.15
0.000209
723.33
1626.41
1792.09
1769.18
628.97
1297.23
0.000234
2566.80
2354.48
2995.77
3190.81
2534.94
1604.56
0.0025
261.41
429.15
173.21
0
0
0
0.000186
190.52
394.87
233.33
318.56
190.52
190.52
0.001267
504.27
1954.66
0
0
0
0


duration
Fstart
Fend
Fpeak
Fmin
Fmax
bandw
0.001375
0.000383
0.001872
0.000866
0.000254
0.000134
122.47
150.76
604.43
424.26
388.29
252.98
295.80
211.16
328.63
718.51
194.15
336.15
200
0
748.78
1237.44
294.83
252.98
234.52
422.12
507.94
360.56
598.72
0
1603.12
320.51
1089.34
1552.01
388.29
803.74
1700
554.32
862.17
1253.99
736.94
803.74
V.sinensis
Ia io
M.macrodactylus
R.ferrumequinum
H.armiger
R.pusillus


15
10
5
0
−5
−10
Principal Component 2
−20
−10
0
10
20
30
Principal Component 1
Bat species
Hipposideros armiger
Myotis macrodactylus
Vespertilio sinensis
Rhinolophus ferrumequinum
Ia io
Rhinolophus pusillus



originally recorded
Overlapping calls
Training
Separation
Echolocation calls
log spectral magnitude
originally recorded
Comminication calls
Feature Extration
Create Mixture
BLSTM model
Echolocation calls
separated
Comminication calls
Model
input
4 BLSTM layers
output


Vespertilio sinensis
train loss
validation loss
Loss (log)
Rhinolophus ferrumequinum
Loss (log)
Hipposideros armiger
Loss (log)
Myotis macrodactylus
Loss (log)
Ia io
Loss (log)
Rhinolophus pusillus
Loss (log)
Epochs