Improving acoustic monitoring of biodiversity using deep learning-based source separation algorithms

Tzu-Hao Harry Lin

Department of Marine Biodiversity Research, JAMSTEC

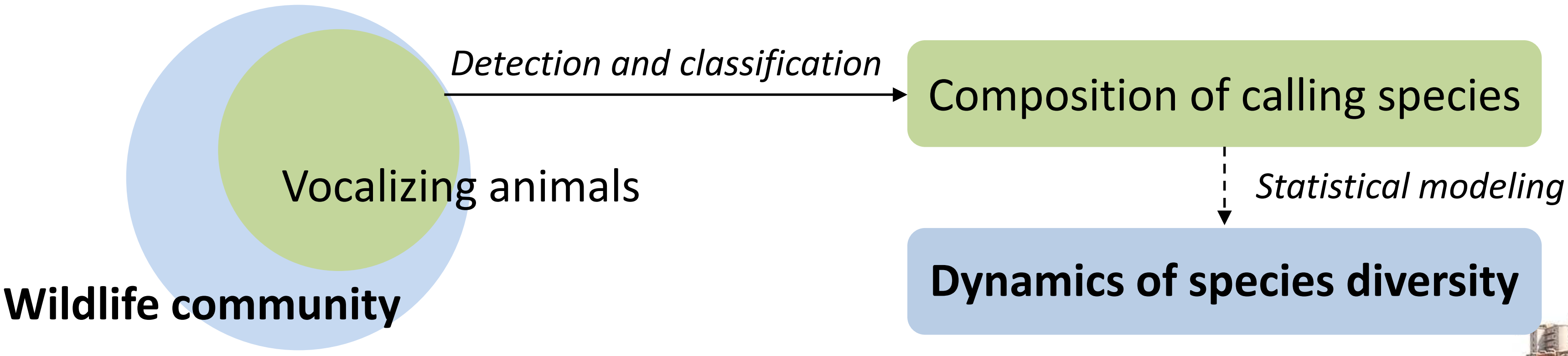
Mao-Ning Tuanmu, Joe Chun-Chia Huang, Chia-Yun Lee

Biodiversity Research Center, Academia Sinica

Yu Tsao

Research Center for Information Technology Innovation, Academia Sinica

Passive acoustic monitoring



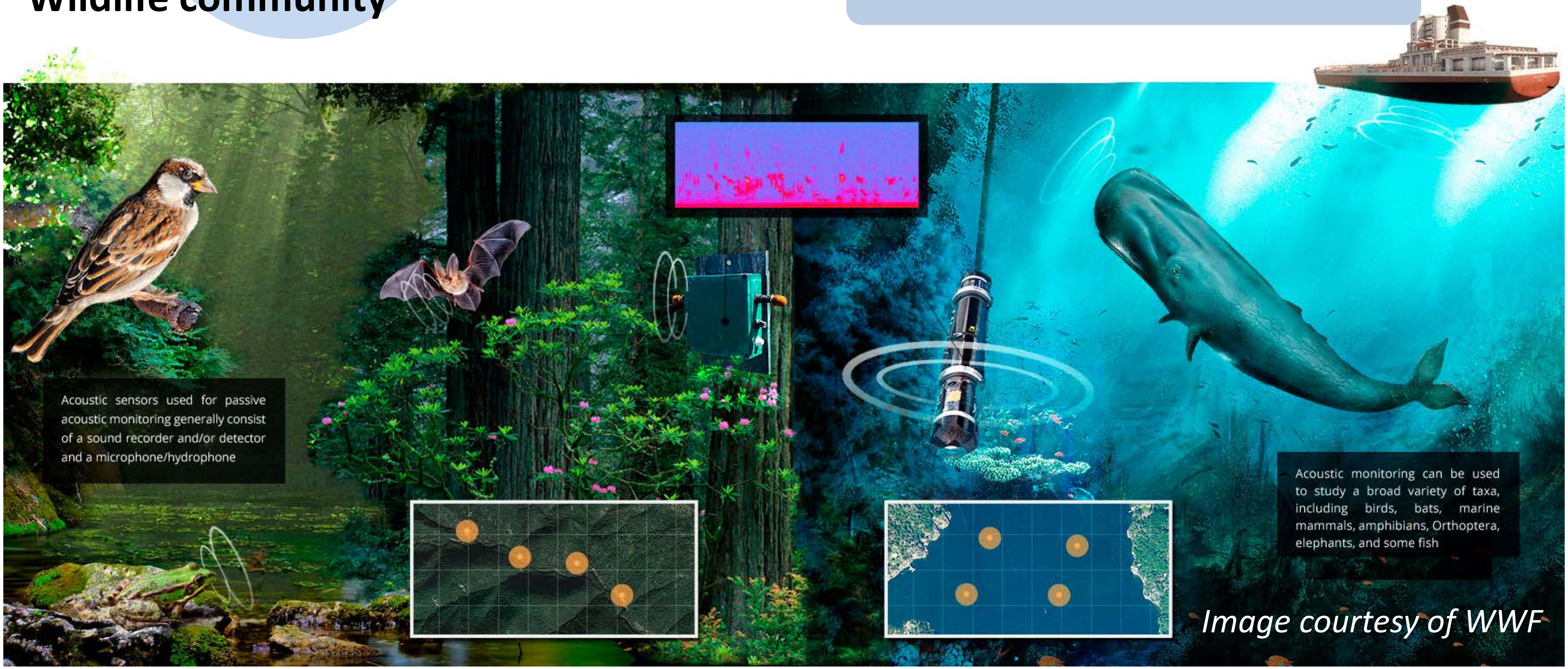
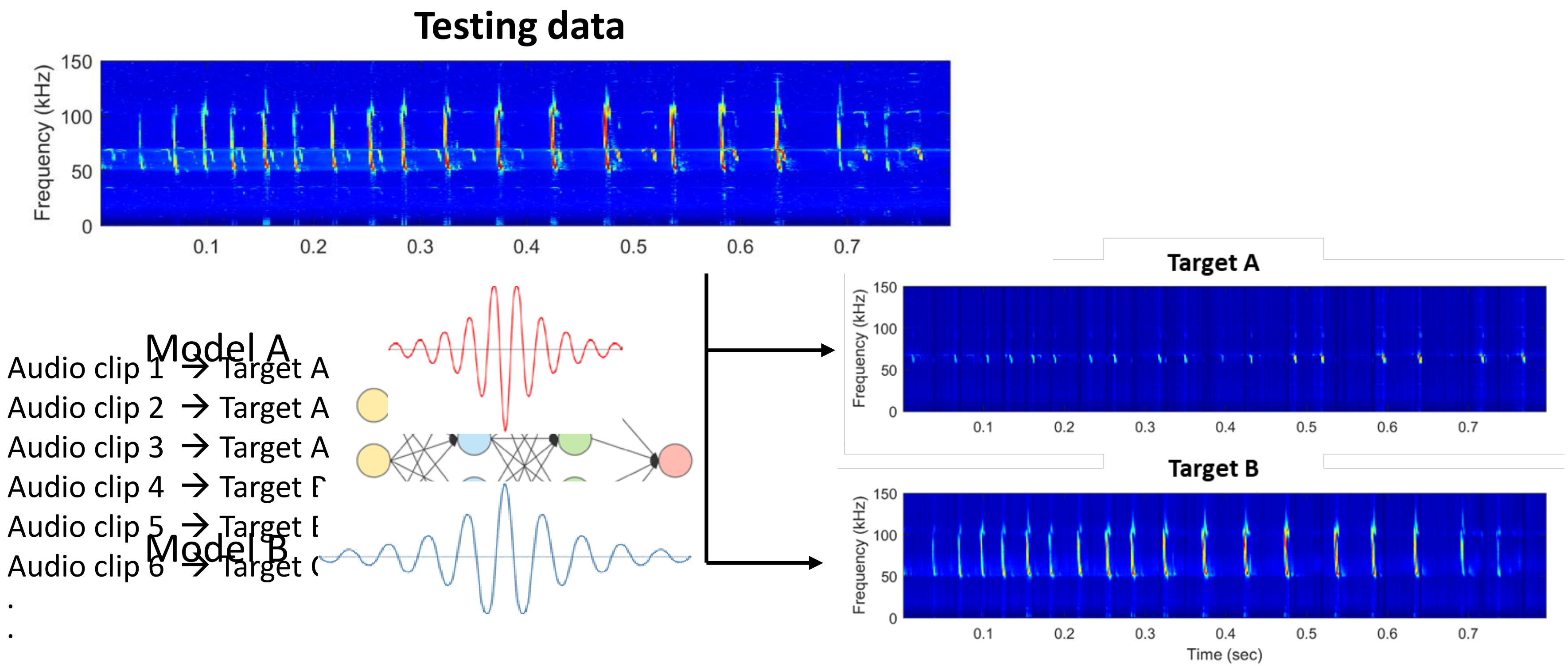


Image courtesy of WWF

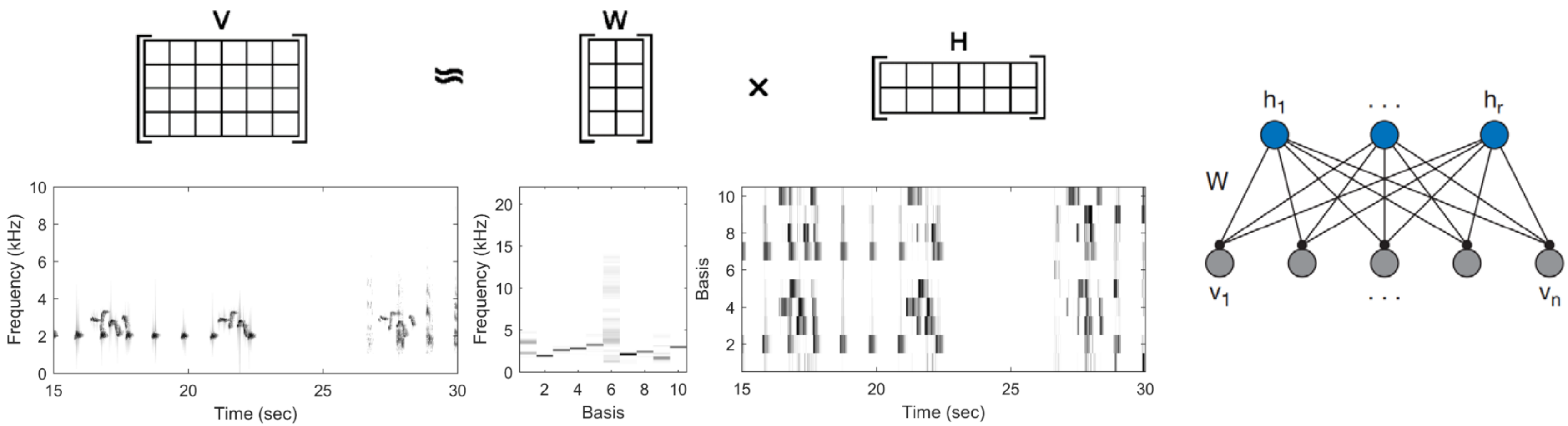
Acoustic information retrieval



- Acoustic information retrieval remains difficult due to lack of training data
- Separate different types of animal vocalizations without a comprehensive training database?

Nonnegative matrix factorization (NMF)

- Decomposition of a spectrogram (castle)
  - **Basis matrix (W)** : spectral features (bricks)
  - **Encoding matrix (H)** : temporal strength (number of bricks)

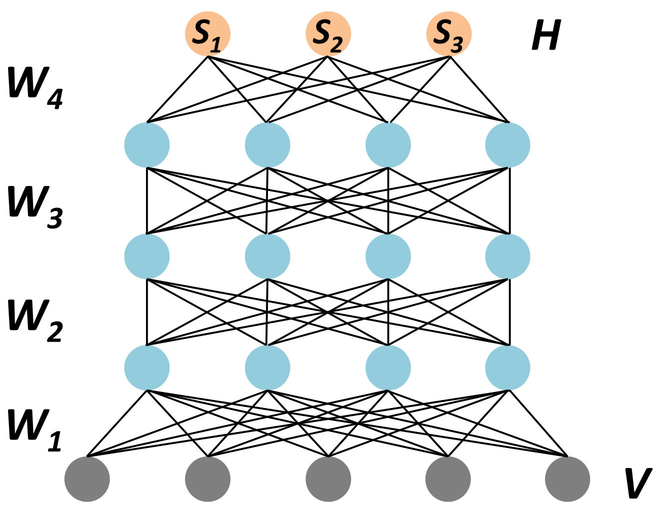


- How to find the appropriate basis cluster for each source?

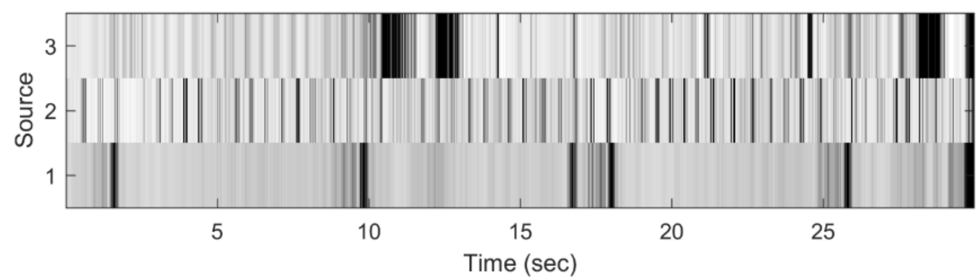
Find the basis weights by multi-layer NMF

- Learn temporal-spectral interactions by a series of convolutive layers
- Learn the encoding information of *k* sources by multi-layers of NMF without any labeled data

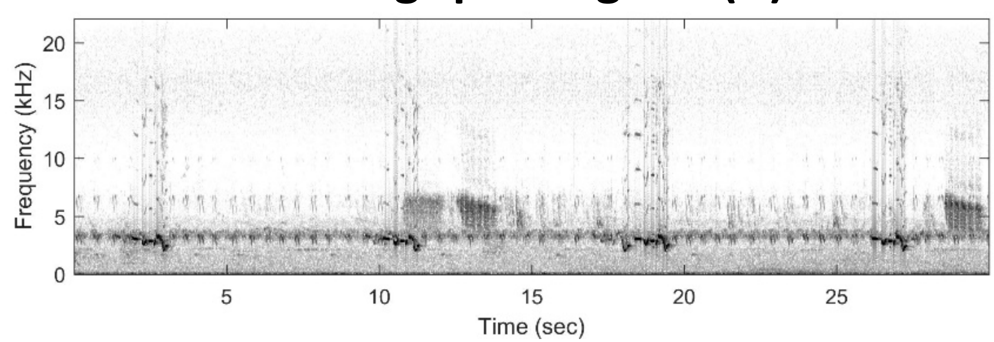
$$V \approx W_1 \cdot W_2 \cdots W_M \cdot H$$



Encoding matrix (H)



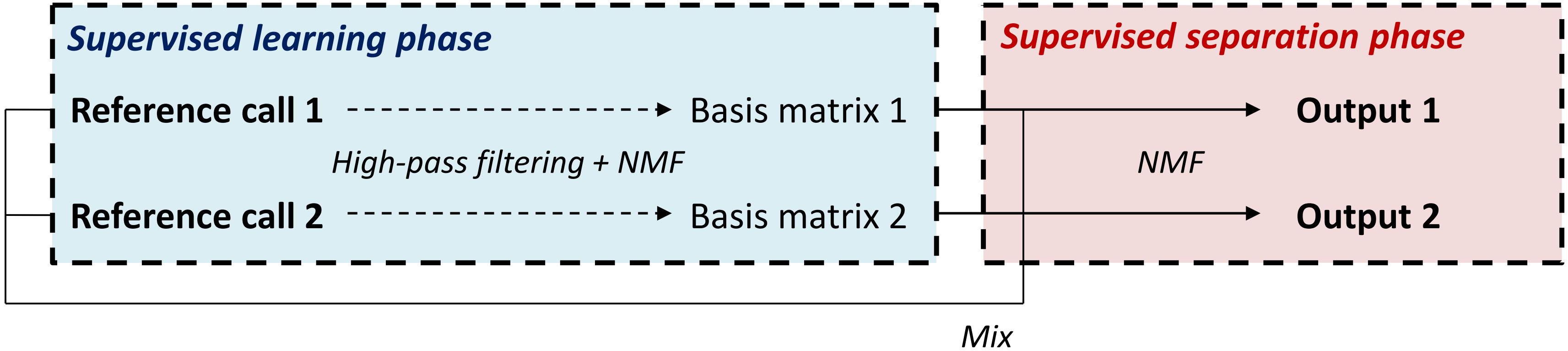
Testing spectrogram (V)



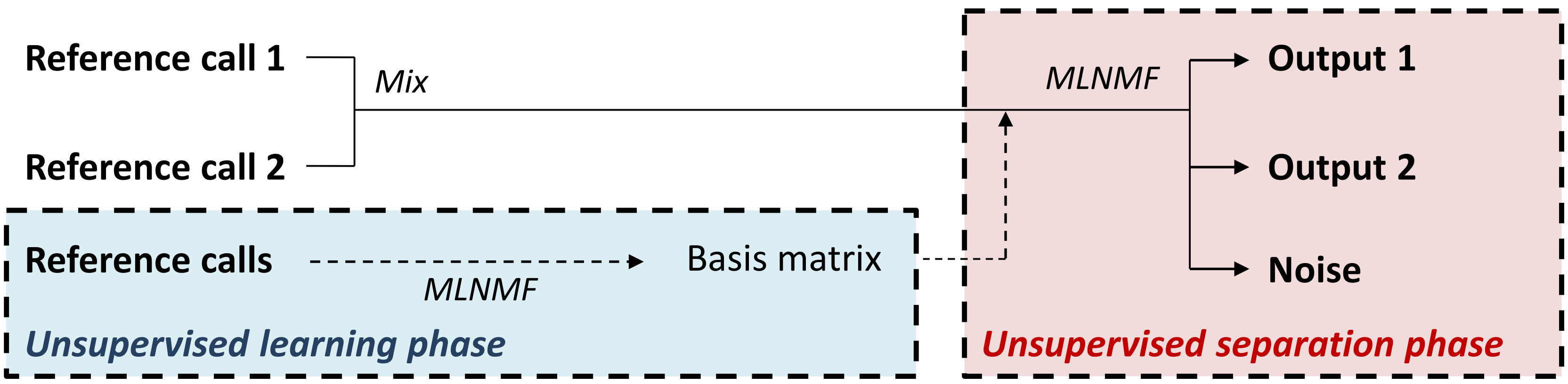
Task 1: separate overlapping calls from simulation data

- Test whether the unsupervised NMF can outperform the supervised NMF

Supervised NMF: NMF

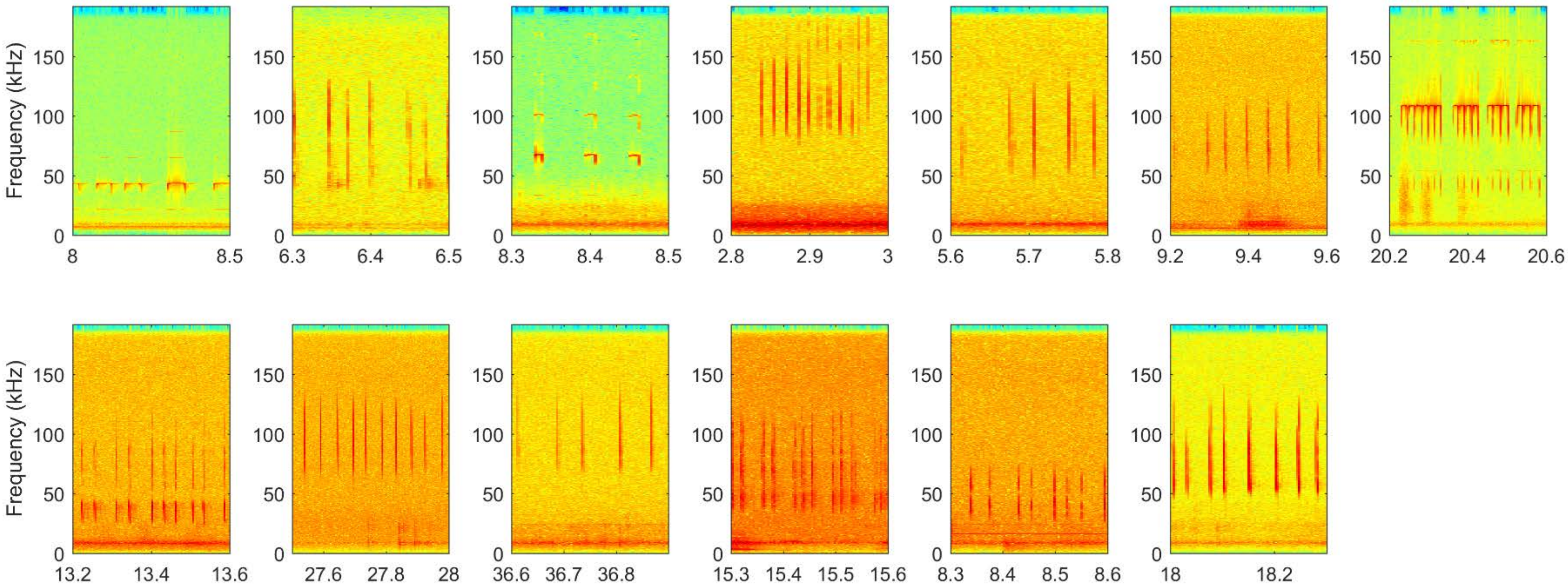


MLNMF: NMF – CNMF (2) – CNMF (4) – CNMF (8)

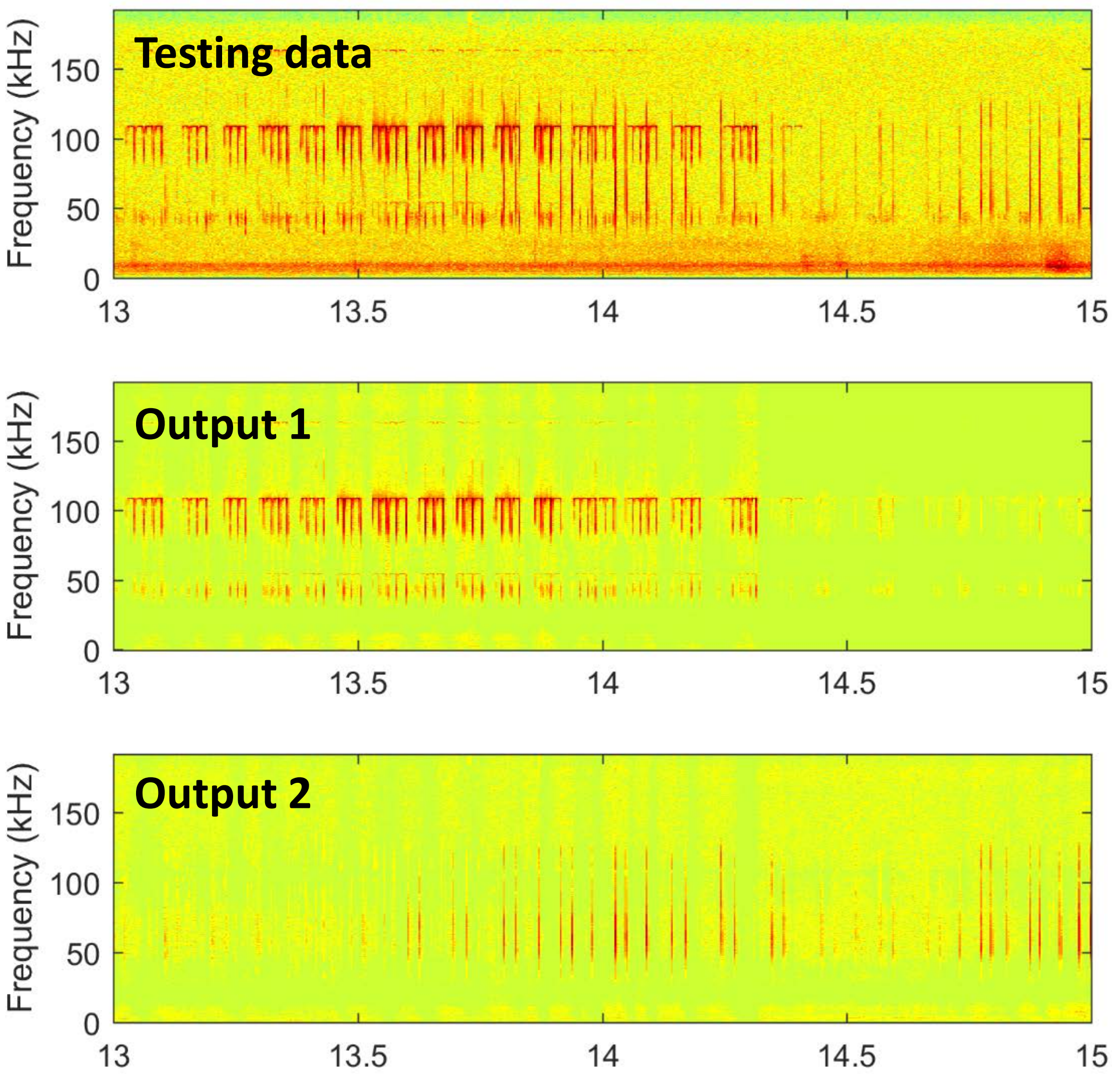


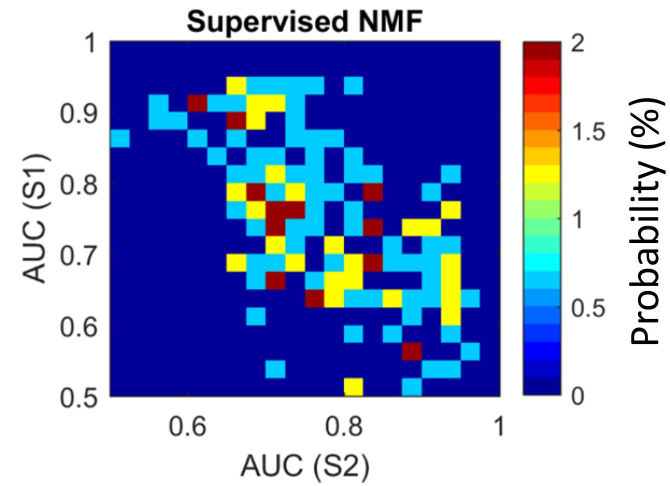
Recordings of reference calls

- Echolocation calls of 13 bat species recorded in Taiwan
  - SNRs are good, but still contain noise
  - Sampling rate: 384 kHz; FFT analysis: 2048 samples, 25% overlap

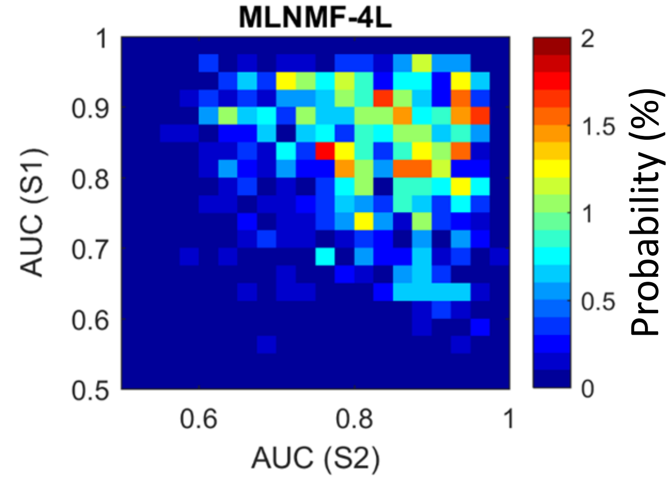


NMF-based BSS outperformed supervised NMF





Mean ± SD: 0.69 ± 0.22

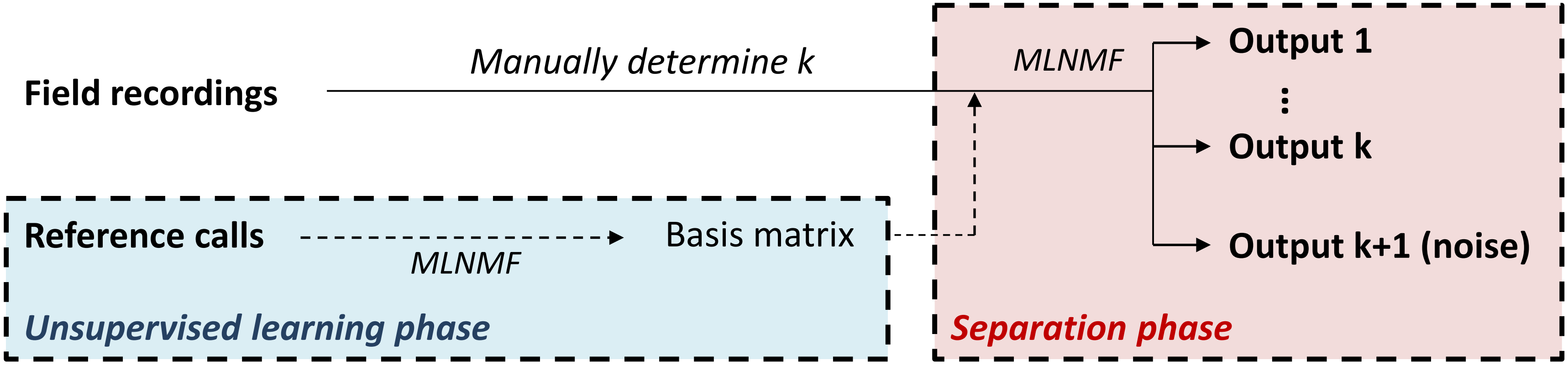


Mean ± SD: 0.77 ± 0.24

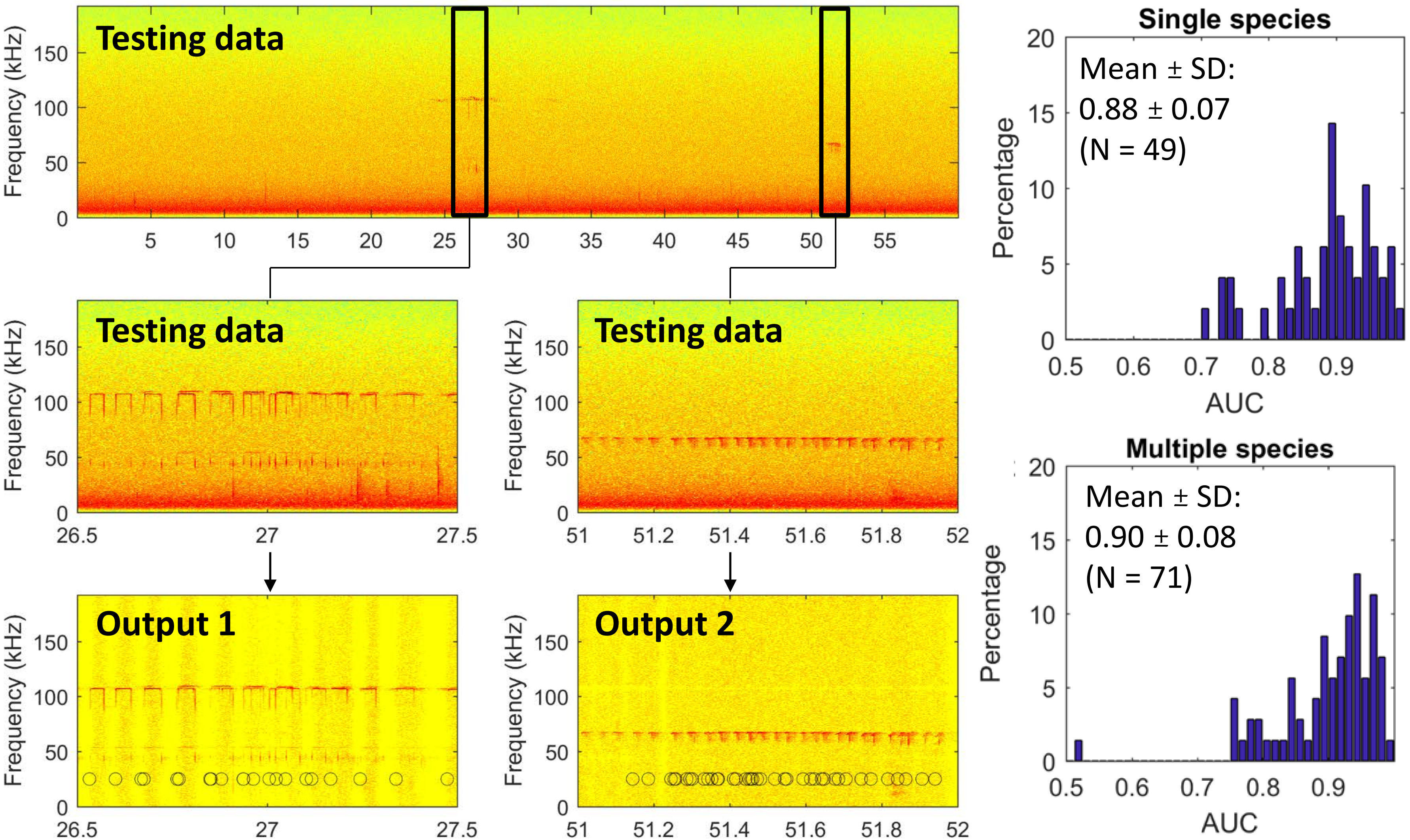
Task 2: separate bat echolocation calls from field data

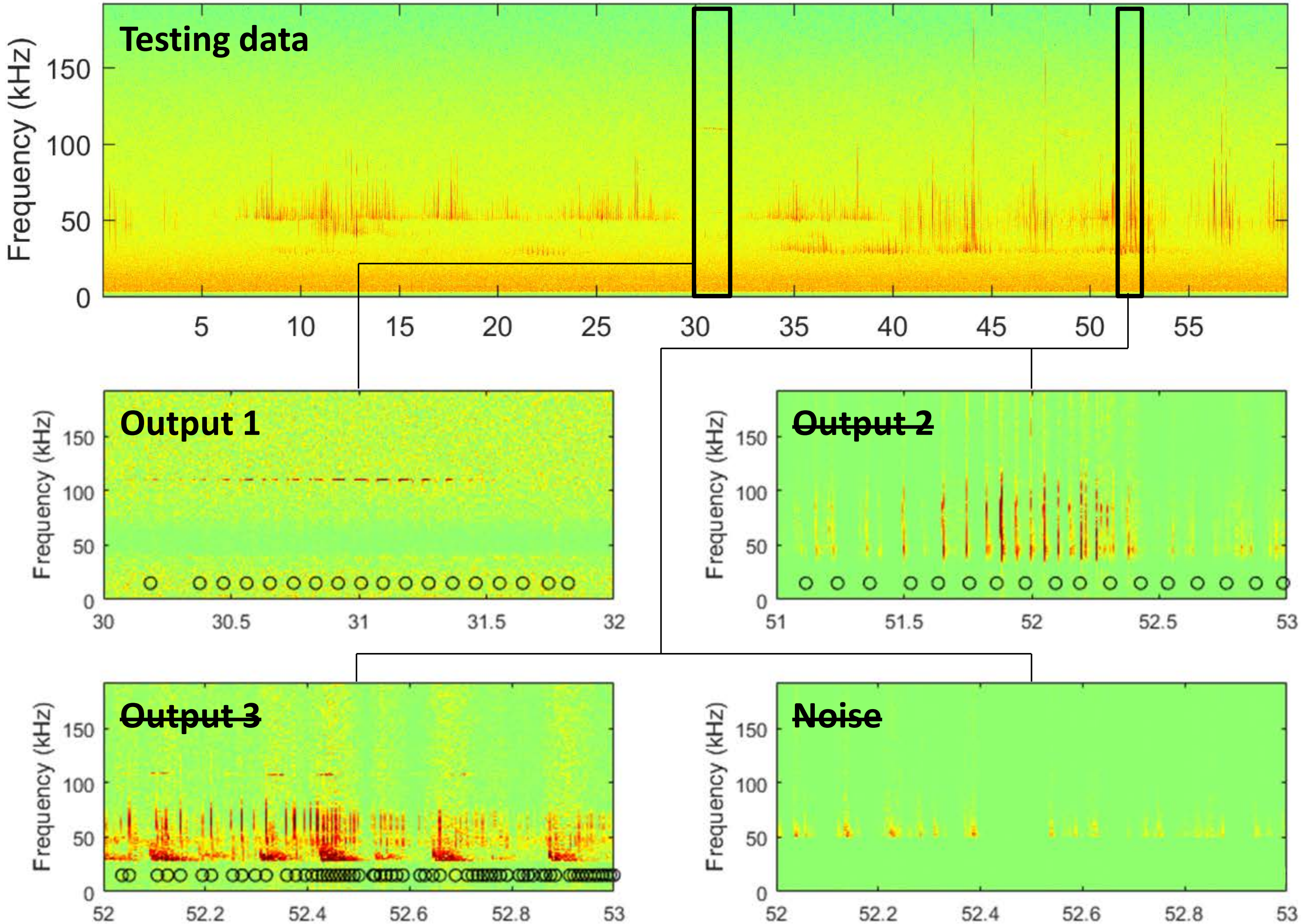
- Test the performance of MLNMF in real-world data
- Field recordings collected in central Taiwan by SM4BAT
  - Sampling rate: 384 kHz; FFT analysis: 2048 samples, 25% overlap
  - 120 one-minute recording clips with various noise conditions

MLNMF: NMF – CNMF (2) – CNMF (4) – CNMF (8)

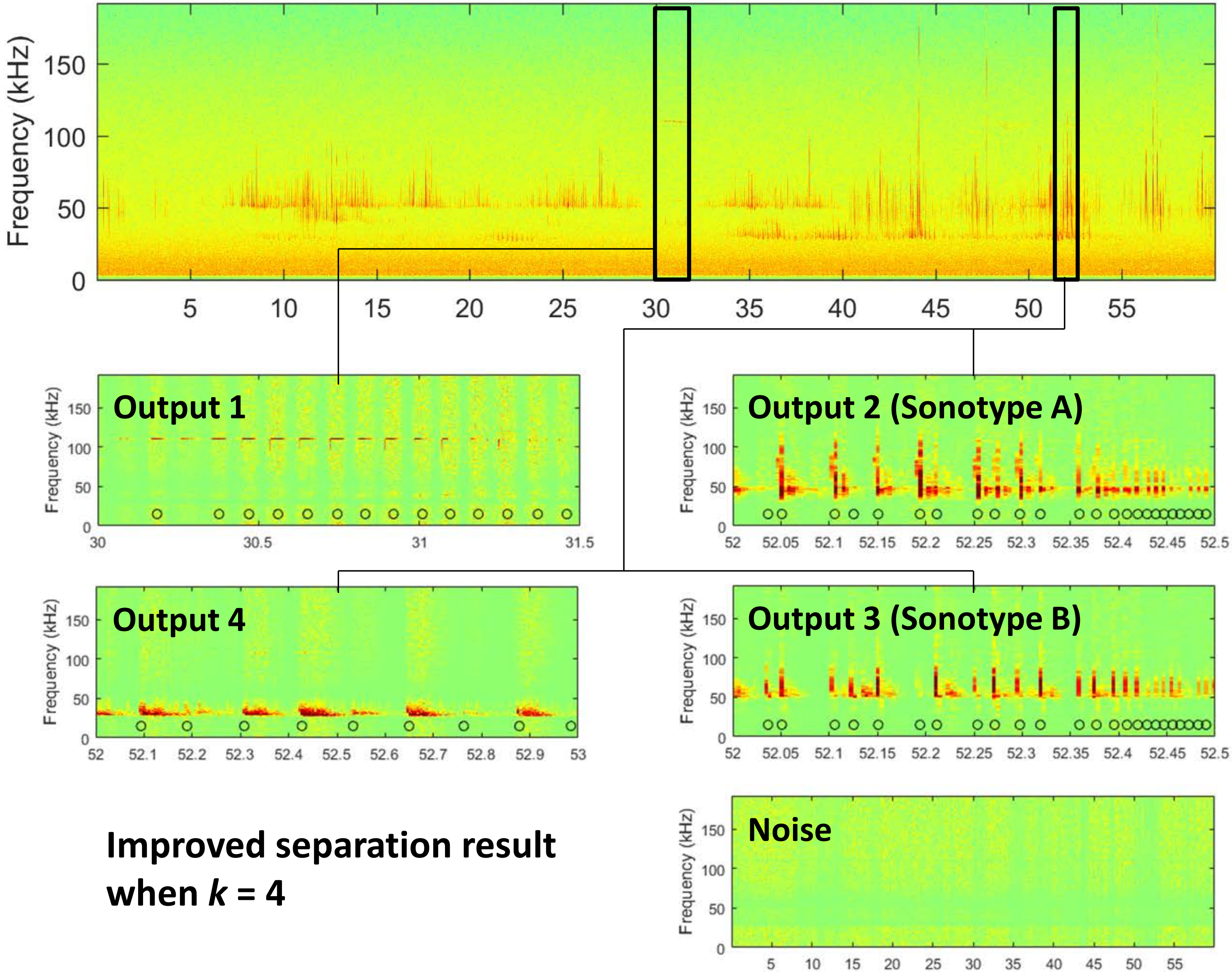


NMF-based BSS can identify most of manually annotated calls





Improper separation result when *k* = 3



Frequency (kHz)
150
100
50
0
5
10
15
20
25
30
35
40
45
50
55
Output 1
Output 2 (Sonotype A)
Output 3 (Sonotype B)
Output 4
Noise
30
30.5
31
31.5
52
52.05
52.1
52.15
52.2
52.25
52.3
52.35
52.4
52.45
52.5
52.6
52.7
52.8
52.9
53
Improved separation result
when k = 4

Conclusions

- **Echolocation calls of different bat species can be separated by multi-layers NMF in an unsupervised manner**
  - Species-specific enhancement of bat echolocation calls
  - Facilitate the establishment of audio recognition database
  - Output is not just a label, it is possible to study the vocal behavior by the separation results

- **Future development:**
  - How to determine *k* sound sources without human intervention?
  - Is it possible to evaluate the repertoire diversity by using the deep learning-based BSS?

Acknowledgement

- Academia Sinica Data Science Seed Grant Program
- Ministry of Science and Technology
  - Independent postdoctoral research fellowship
- Asi@Connect Project & TEIN*CC

Thanks for your attention!