Improving biodiversity monitoring through soundscape information retrieval

Yu Tsao, Tzu-Hao Lin

Research Center for Information Technology Innovation, Academia Sinica

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

Biodiversity Research Center, Academia Sinica

Chiou-Ju Yao

National Museum of Natural Science

Dr. Yu Tsao (曹昱), *Asst. Research Fellow*

- Education

  - Ph.D. in ECE, Georgia Institute of Technology, 2003-2008
  - M.S. in EE, National Taiwan University, 1999-2001
  - B.S. in EE, National Taiwan University, 1995-1999

- Work Experience

  - Researcher, National Institute of Information and Communications Technology, Spoken Language Communication Group, Japan (2009/4-2011/9)
  - Summer Research Associate, Texas Instruments Incorporated, Speech Technologies Laboratory DSP Solutions R&D Center, United States (2004, 2005, 2006 summers)

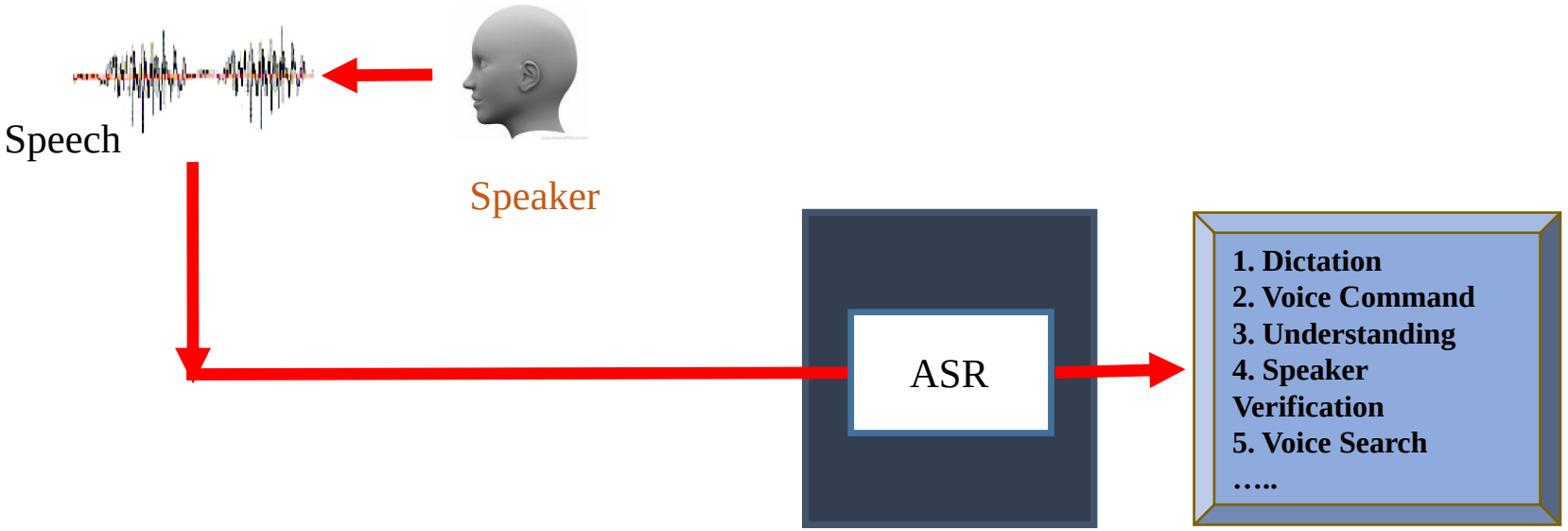
- Research Interests

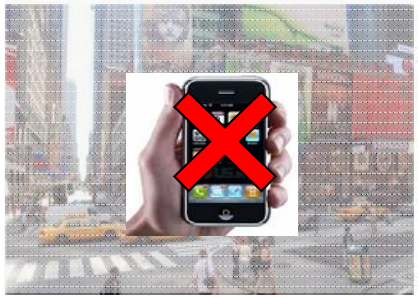
  Speech & Audio Signal Processing, Machine Learning and Pattern Recognition, Speech and Speaker Recognition

- Lab at CITI (Academia Sinica)

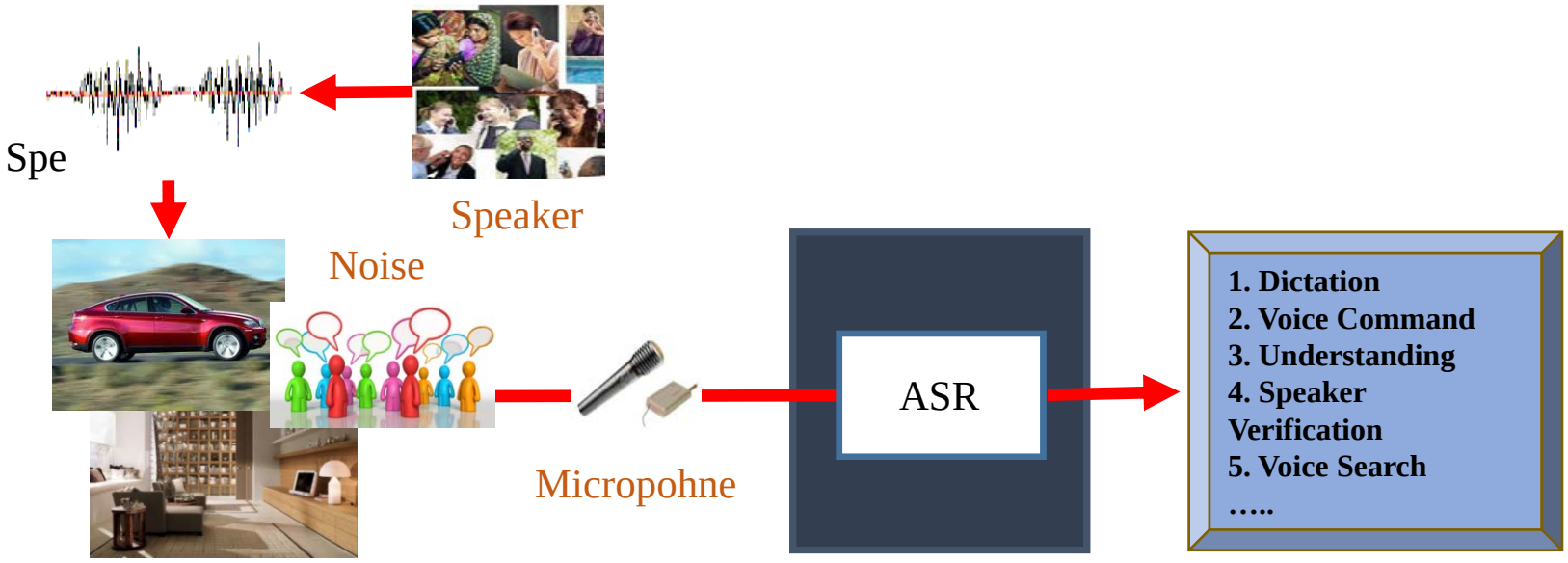
  Bio-Acoustic Signal Processing (Bio-ASP) Lab

Automatic Speech Recognition





Automatic Speech Recognition



- Mismatch issues may come from
  - Speaker (age, status, gender)
  - Environment (distance, microphone, noise)
    - All of these issues may happen together

自動語音辨識 (Automatic Speech Recognition)

語音辨識度 (字辨識率 in %)

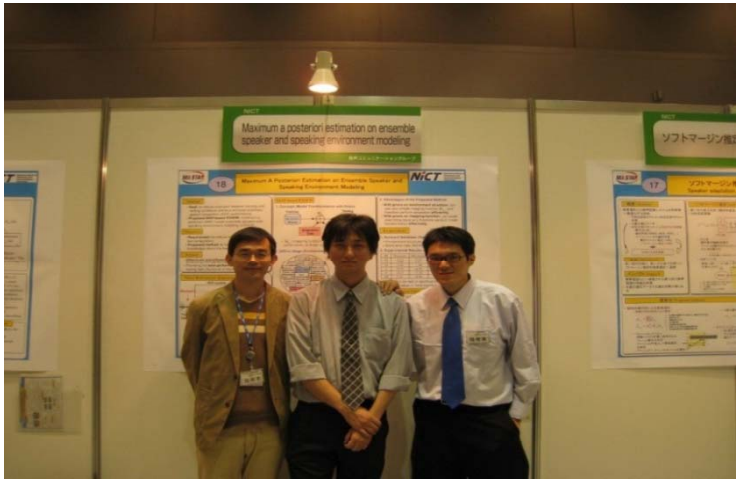
SNR (dB)	Clean	20	15	10	5	0	Avg
Baseline	99.64	95.09	85.59	65.14	36.51	14.85	59.44
MMSE	99.61	96.38	90.38	77.3	52.72	24.62	68.28
MLSA	99.66	95.37	87.58	70.87	43.95	17.81	63.12
MAPA	99.64	96.45	90.44	77.11	52.25	24.21	68.09
GMAPA	**99.67**	**96.97**	**92.28**	**80.38**	**56.87**	**27.79**	**70.86**

所提出的訊號處理、模式識別技術使辨識率**可以有效地提升** Baseline 系統 ，字錯誤率降低了 28% (59.44% to 70.86%)

National Institute of Information and Communications Technology (NICT)

Located in Kansai Science City or “Keihanna” area

National Institute of Information and Communications Technology (NICT)



Speech Translation System

- Speech-to-speech translation (voice)
- Speech-to-speech translation (text)

Speech Translation on IPhone (VoiceTra4U)

iPhone 音声翻訳

NiCT 独立行政法人 情報通信研究機構
National Institute of Information and Communications Technology

- Speech-to-speech translation (voice)

哆啦A夢科學館 http://www.doraemon-miraiten.jp/intro/index.html

U-STAR Project

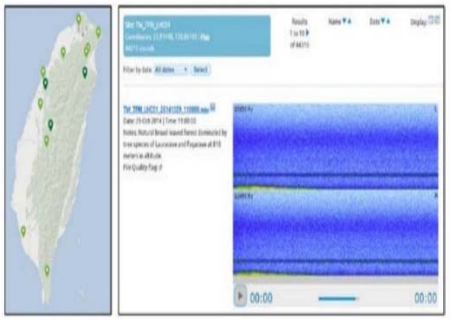
http://www.ustar-consortium.com/members.html

Speech-related Applications

Education

Chatbot

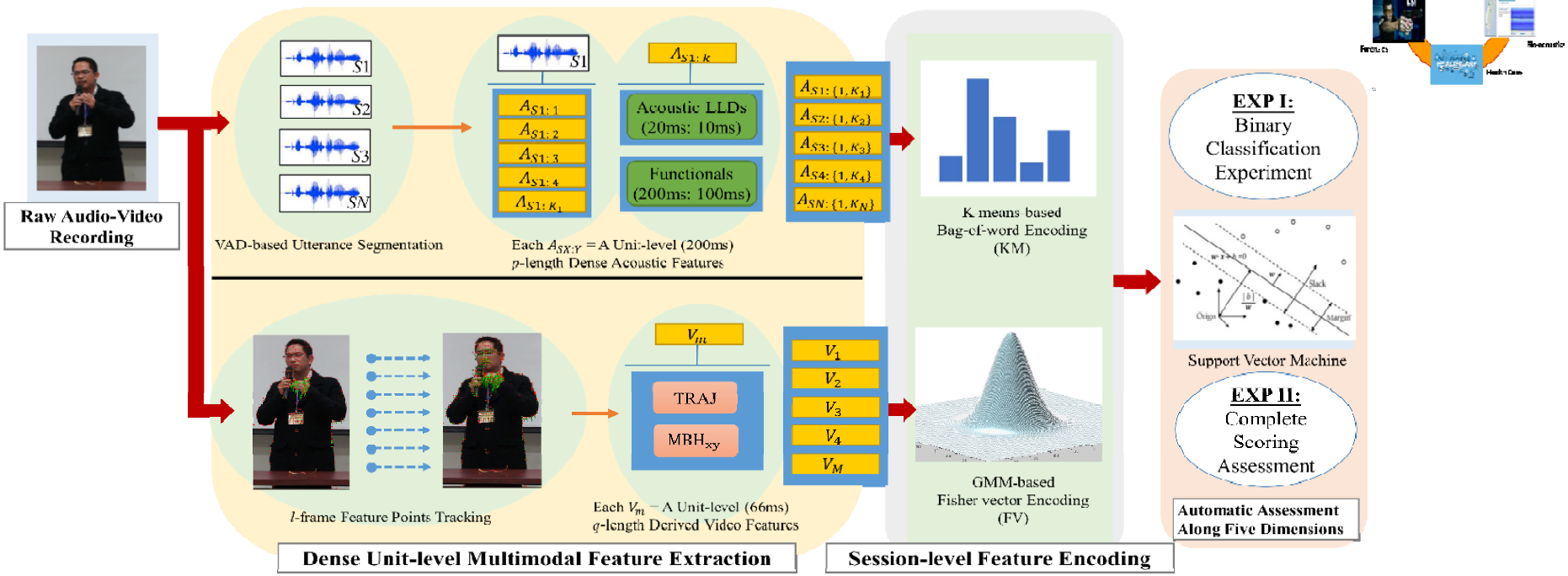
Forensics



Bio-acoustics

Health Care

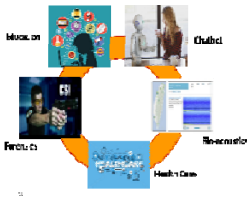
Education (Presentation Training)

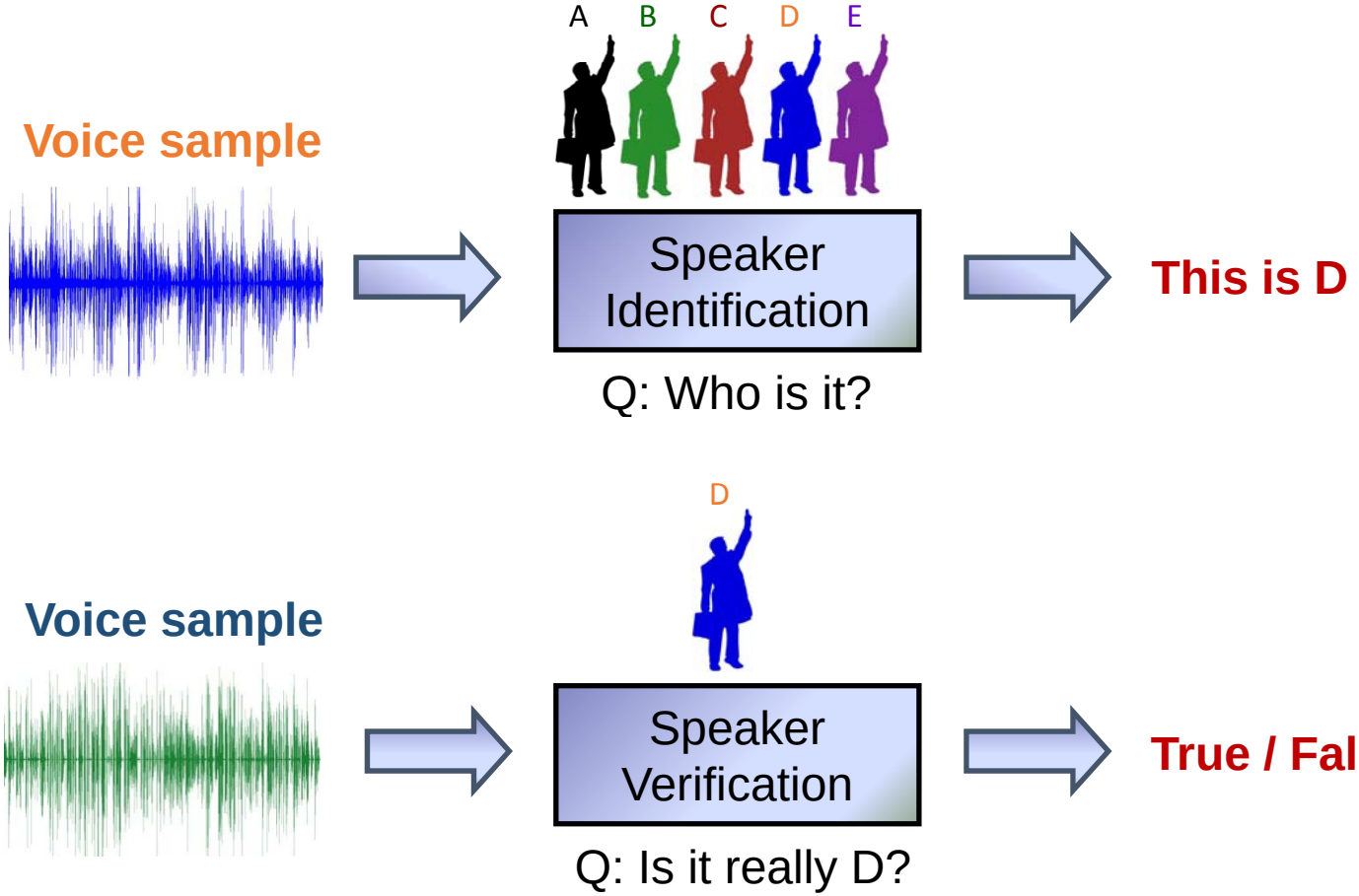


	Baseline SVR			Binary-SVR			Proposed Binary-SVM		
	Audio	Video	Multimodal	Audio	Video	Multimodal	Audio	Video	Multimodal
Dim$_{1o}$	0.316	0.398	0.445	0.382	0.423	0.495	0.391	0.478	**0.534**
Dim$_{2o}$	0.227	0.380	0.394	0.252	0.385	0.411	0.195	0.397	**0.421**
Dim$_{3o}$	0.324	0.457	0.482	0.349	0.482	0.519	0.346	0.511	**0.552***
Dim$_{1r}$	0.300	0.330	0.432	0.294	0.332	0.433	0.350	0.360	**0.470**
Dim$_{2r}$	0.304	0.278	0.362	0.300	0.278	0.360	0.310	0.294	**0.417**
Dim$_{3r}$	0.455	0.367	0.539	0.462	0.362	0.542	0.475	0.357	**0.551***

S.-W. Hsiao et al., “Toward Automating Oral Presentation Scoring during Principal Certification Program using Audio-Video Low-level Behavior Profiles”. IEEE TAC, 2017.

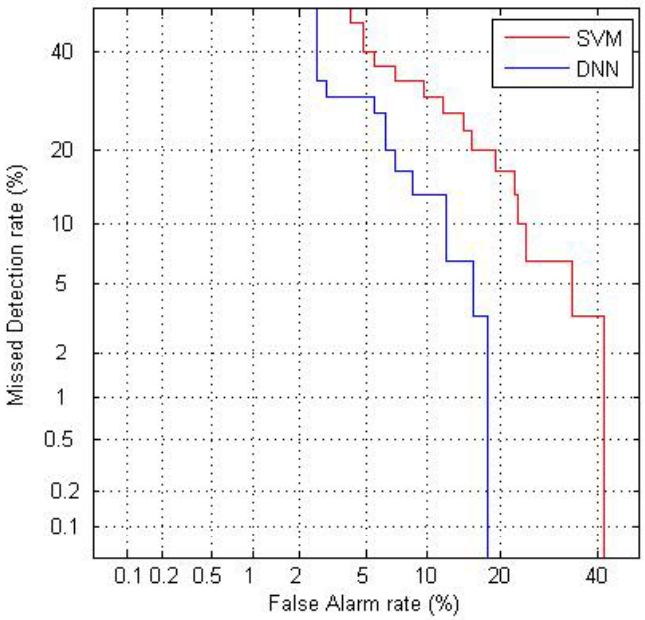
Forensics (Speaker Recognition)





Speaker Verification

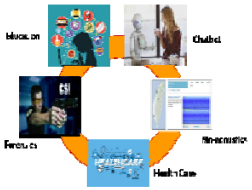
- Detection error tradeoff curve(DET curve)
  - A graphical plot of error rates for binary classification systems, plotting false alarm rate vs. missed detection rate

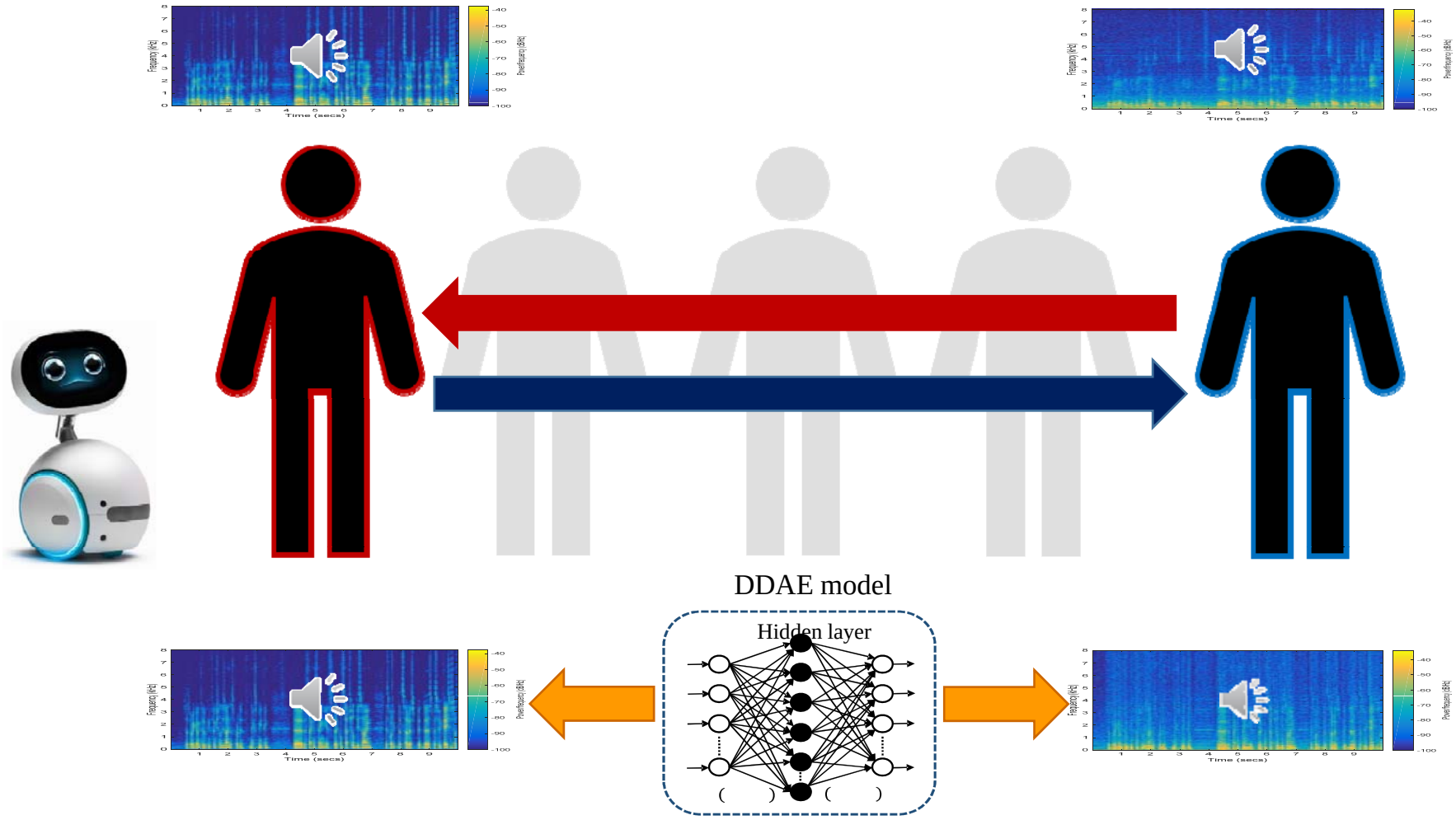


	EER
SVM	19.26%
DNN	12.22%

DNN outperforms SVM when using I-vector as the feature extraction in all of the evaluations.

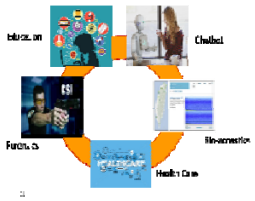
Chatbot (Distant Talking)

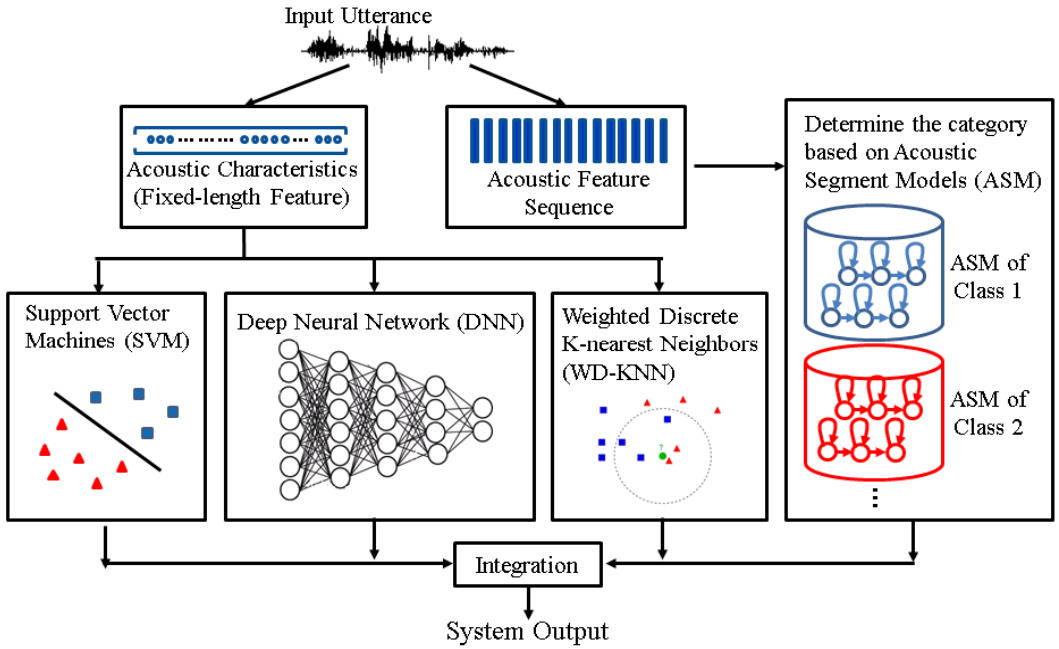




P. Lin et al., “Multi-style Learning with Denoising Autoencoders for Acoustic Modeling in the Internet of Things (IoT)”. CSL, 2017.

Health Care (Speech Emotion Recognition)



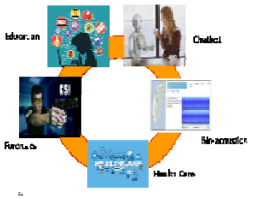


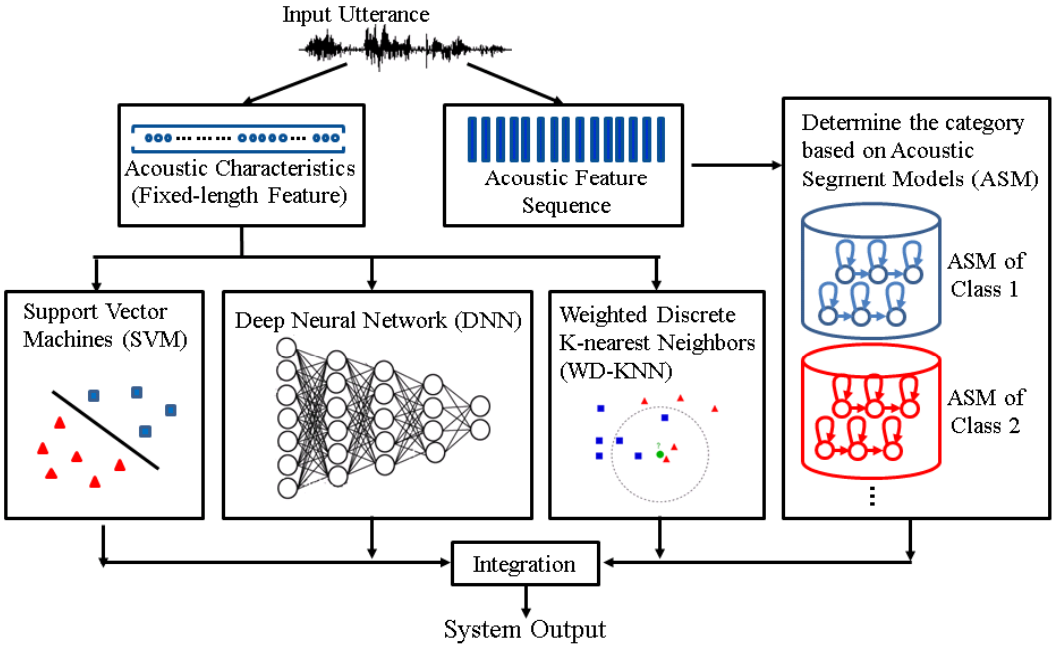
- The proposed ensemble system that incorporates **multiple knowledge sources** outperforms individual systems

UAR*	Arousal	Valence	12-way Emotion
Baseline	82.4%	77.9%	40.1%
Structural SVM	83.2%	76.9%	45.0%
DNN	87.7%	81.2%	47.6%
KNN	82.4%	52.4%	40.1%
ASM	72.9%	66.3%	22.4%
Ensemble	**88.2%**	**84.1%**	**49.4%**

* unweighed average recall

Health Care (Autism Disorders Recognition)



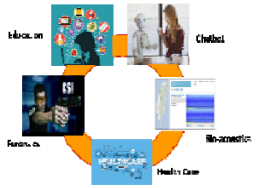


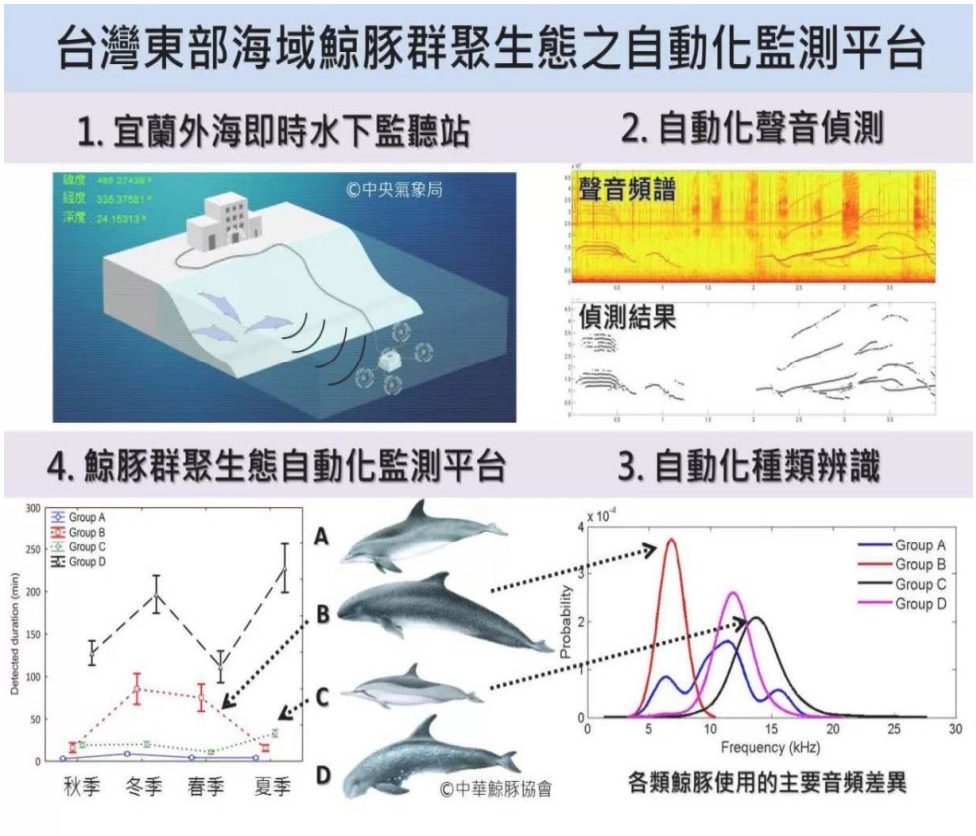
- The proposed ensemble system that incorporates **multiple knowledge sources** outperforms individual systems

UAR	Typical	Diagnosis
Baseline	92.8%	52.4%
Structural SVM	93.2%	56.2%
DNN	94.4%	57.5%
KNN	84.1%	51.7%
ASM	70.2%	32.9%
Ensemble	**94.5%**	**57.8%**

H.-y. Lee et al., "Ensemble of Machine Learning and Acoustic Segment Model Techniques for Speech Emotion and Autism Spectrum Disorders Recognition". Interspeech 2013.

Bio-acoustics (Biology Monitoring)





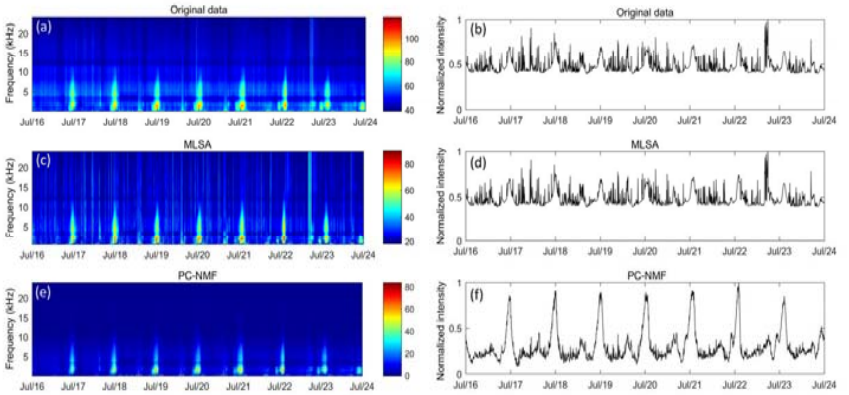
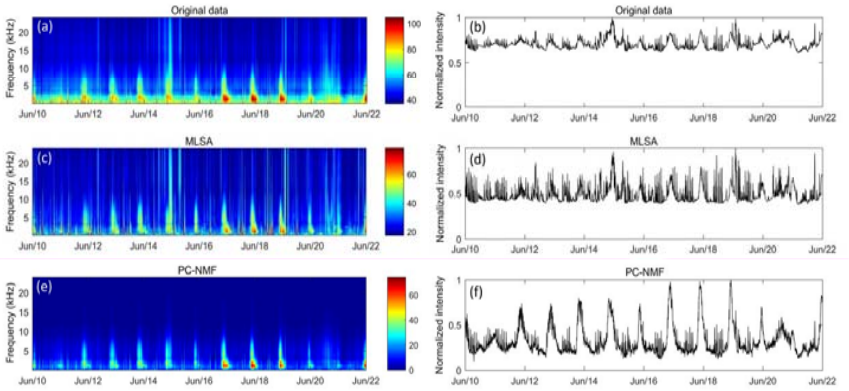


Figure 5. Spectrograms and normalized intensities of underwater recordings collected at the Waisanding Sandbar. The graph shows a comparison between (**a**,**b**) the original data and (**c**,**d**) the analysis results using the MLSA method and (**e**,**f**) the PC-NMF.



T.-H. Lin et al., ""Improving biodiversity assessment via unsupervised separation of biological sounds from long-duration recordings". SR, 2017

Improving biodiversity monitoring through soundscape information retrieval

Yu Tsao, Tzu-Hao Lin

Research Center for Information Technology Innovation, Academia Sinica

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

Biodiversity Research Center, Academia Sinica

Chiou-Ju Yao

National Museum of Natural Science

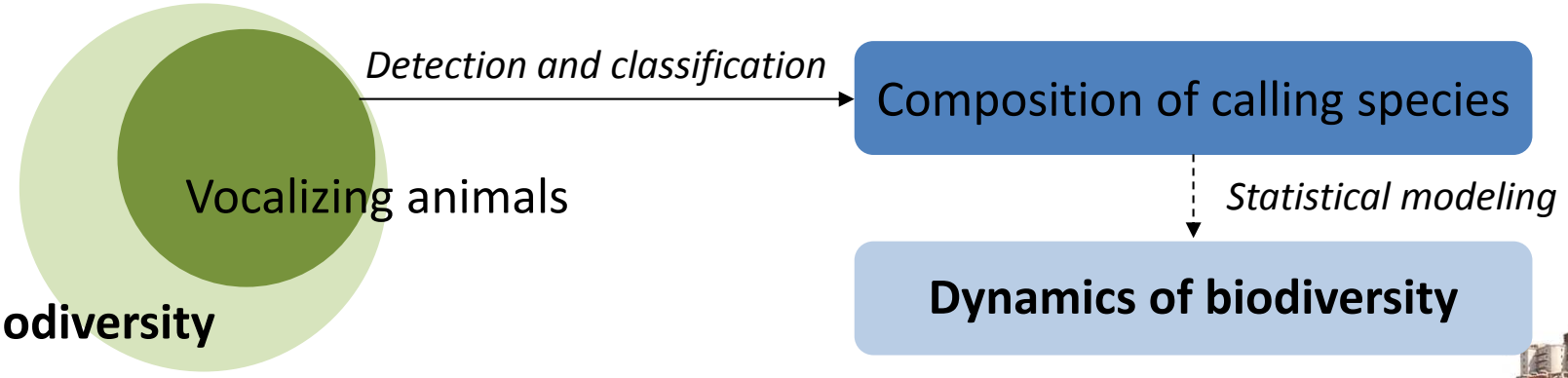
Soundscape: data science of ecosystem

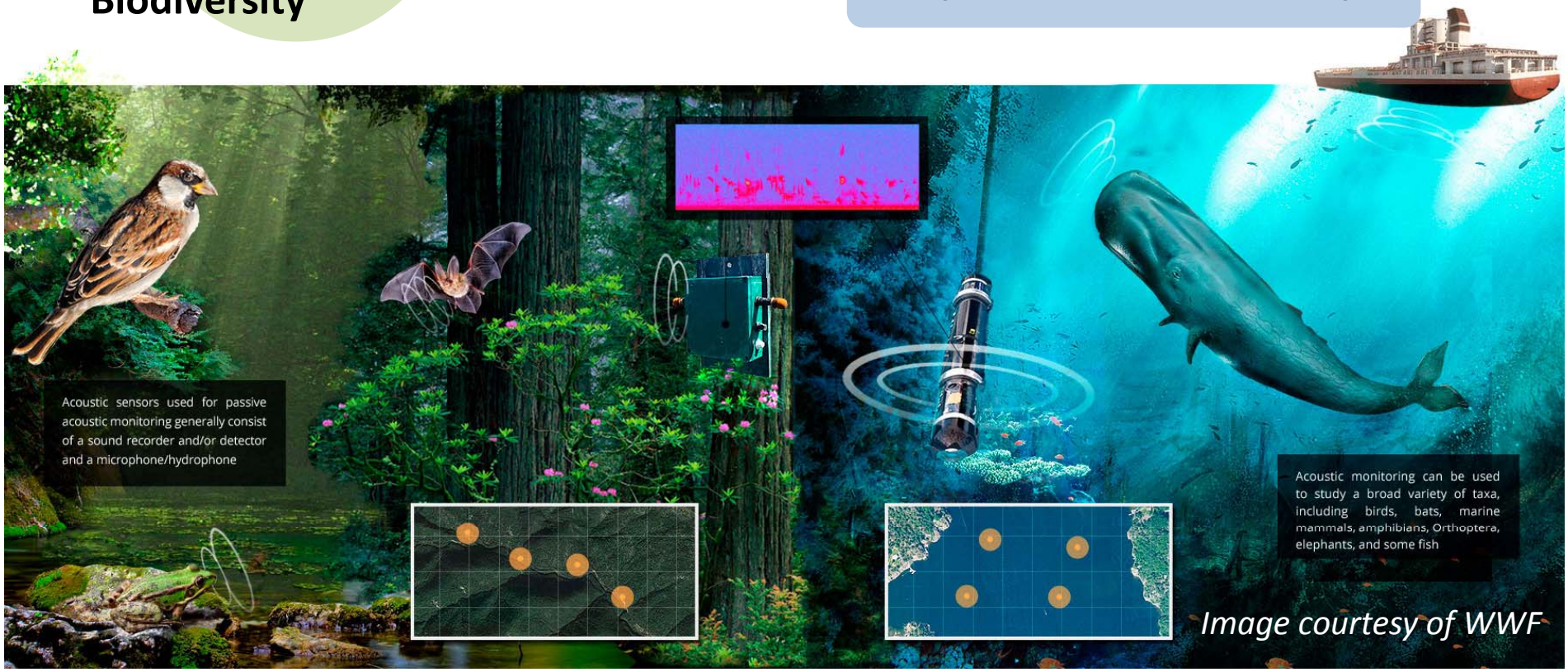
- **Biophony:** biodiversity
- **Geophony:** geophysical events and weather conditions
- **Anthrophony:** impacts of human activities on ecosystem

Large-scale soundscape projects

- **Big data:** large-scale, long-term monitoring
- **Interdisciplinary integration:** sensor network, signal processing, ecological research

Concept of passive acoustics-based biodiversity monitoring

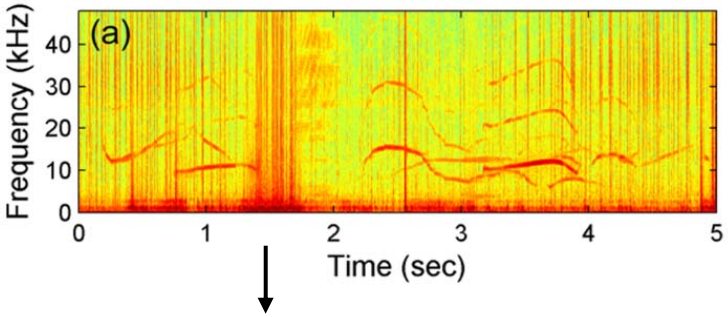




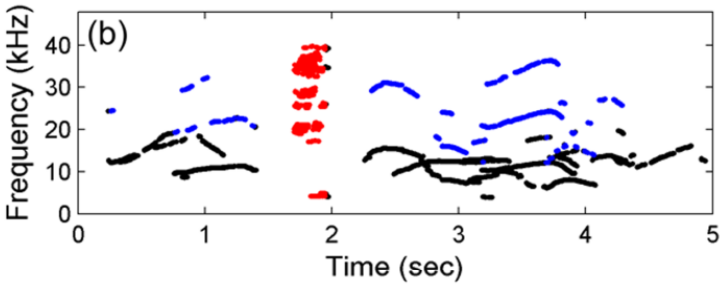
Searching based on existing information

Choose targets → Rule-based detectors → Long-term monitoring

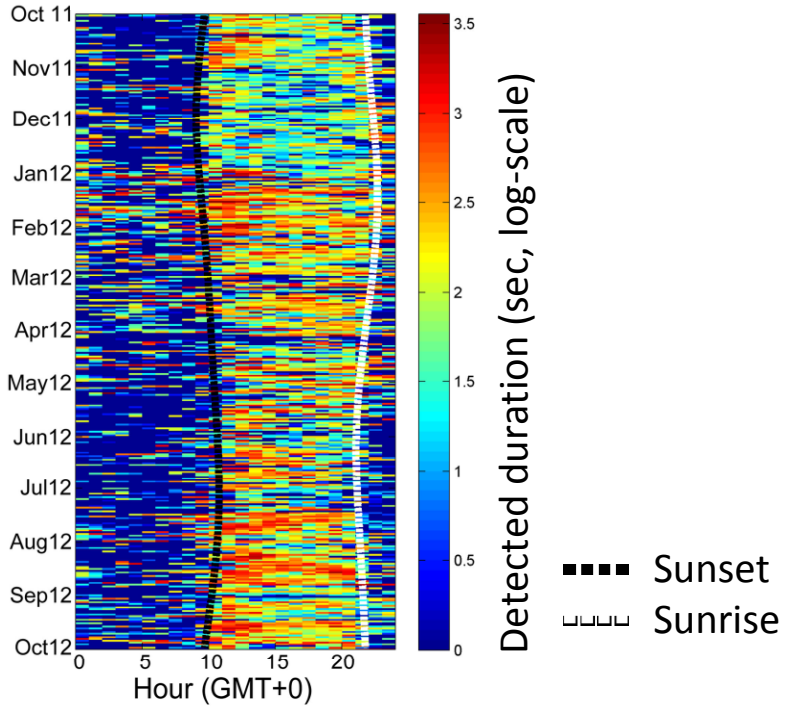
Spectrogram of a recording clip



Detection result of tonal sounds



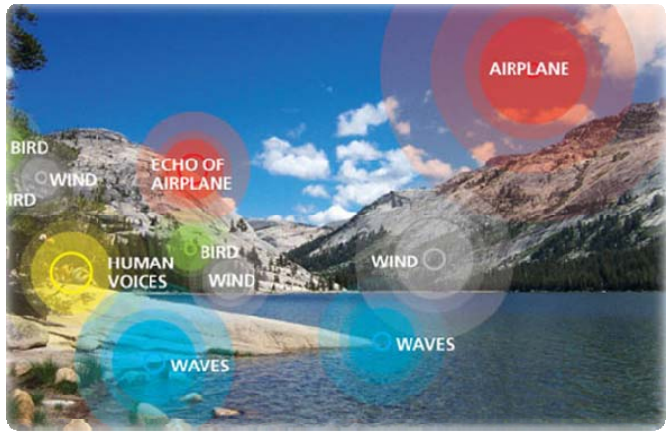
Year-round presence of toothed whales



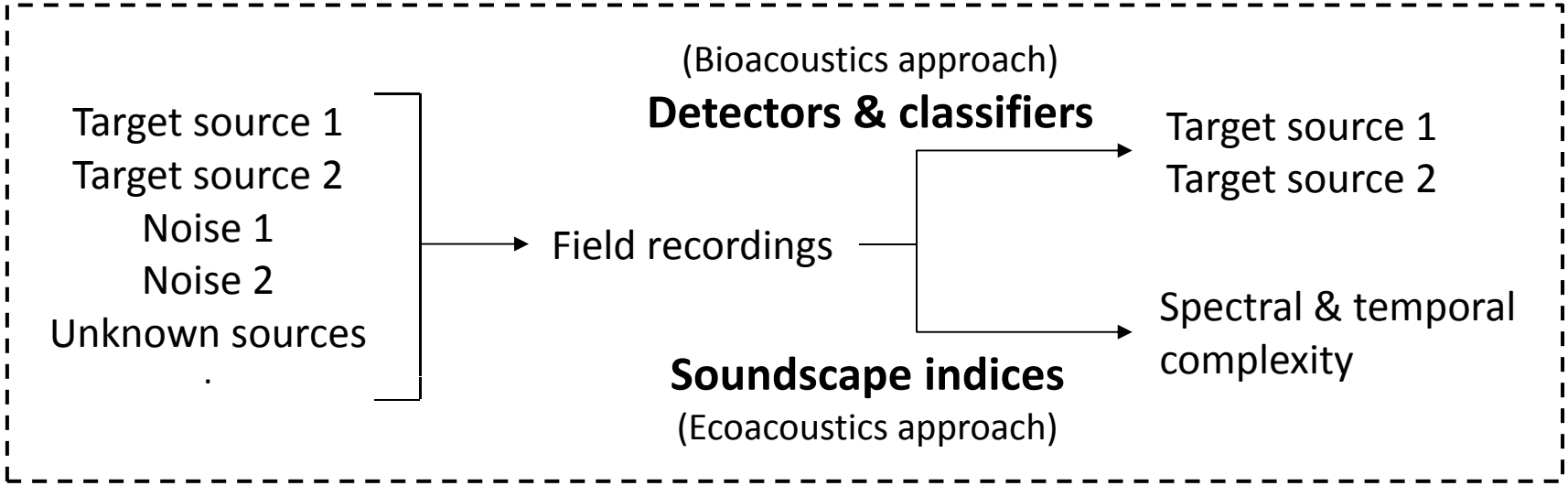
Lin et al. (2015) PLOS ONE

Challenge in soundscape analysis

- **Precision of acoustical analysis**
  - Noise interference
  - Simultaneous sound sources
- **Low precision may lead to a biased ecological interpretation**



https://www.nps.gov/yose/learn/nature/soundscape.htm



Source separation may improve the analysis of soundscape

- **Supervised approach**
  - Model building based on training data
  - Powerful but require labeled and clean data

- **Unsupervised approach (blind source separation)**
  - Dictionary learning
  - Matrix factorization

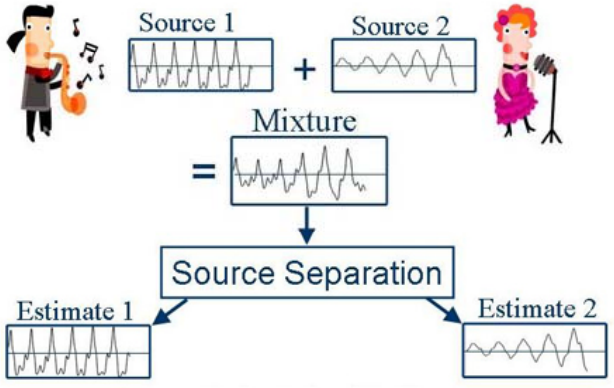
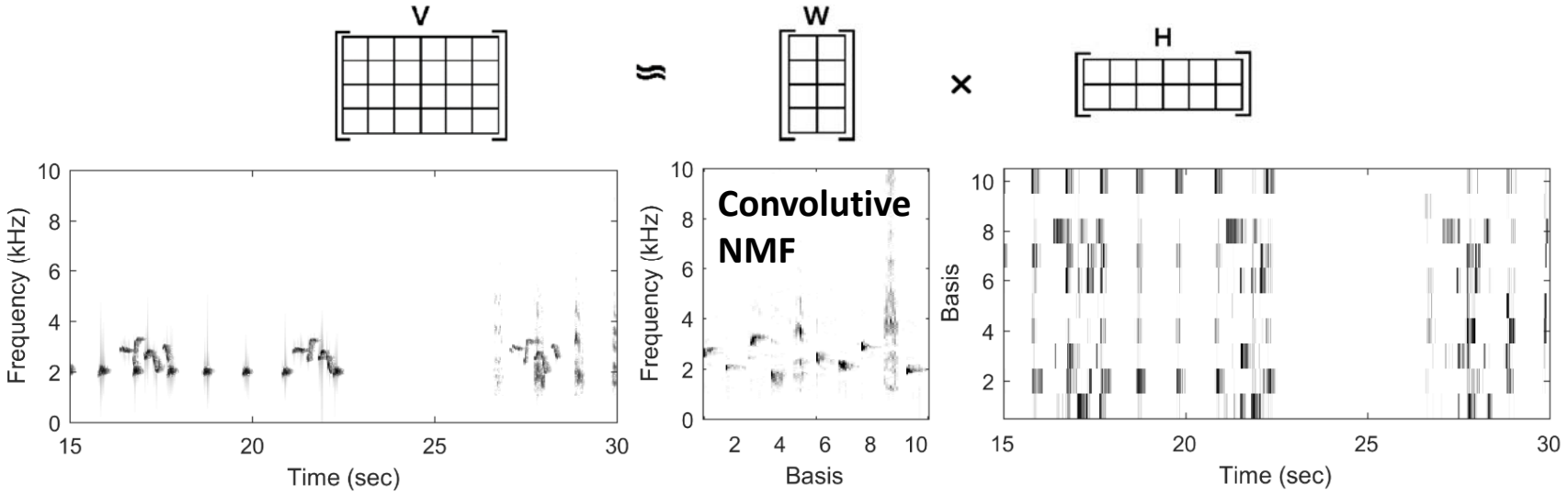


Image courtesy of music.cs.northwestern.edu

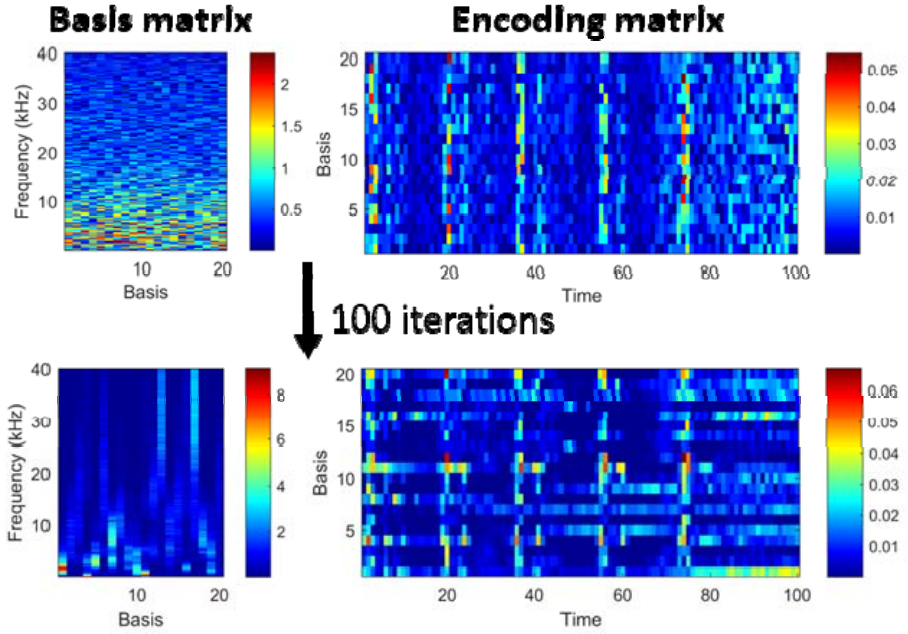
Non-negative matrix factorization (NMF)

- Decomposition of a non-negative matrix (spectrogram)
  - **Basis matrix (W)** : spectral feature (bricks)
  - **Encoding matrix (H)** : temporal information (number of bricks)



Self-learning of NMF

- Initialize by random values or prior knowledge
- Update the basis matrix (*W*) and encoding matrix (*H*) through iterations



Update procedure:

$$W_{ia} \leftarrow W_{ia} \sum_{\mu} \frac{V_{i\mu}}{(WH)_{i\mu}} H_{a\mu}$$

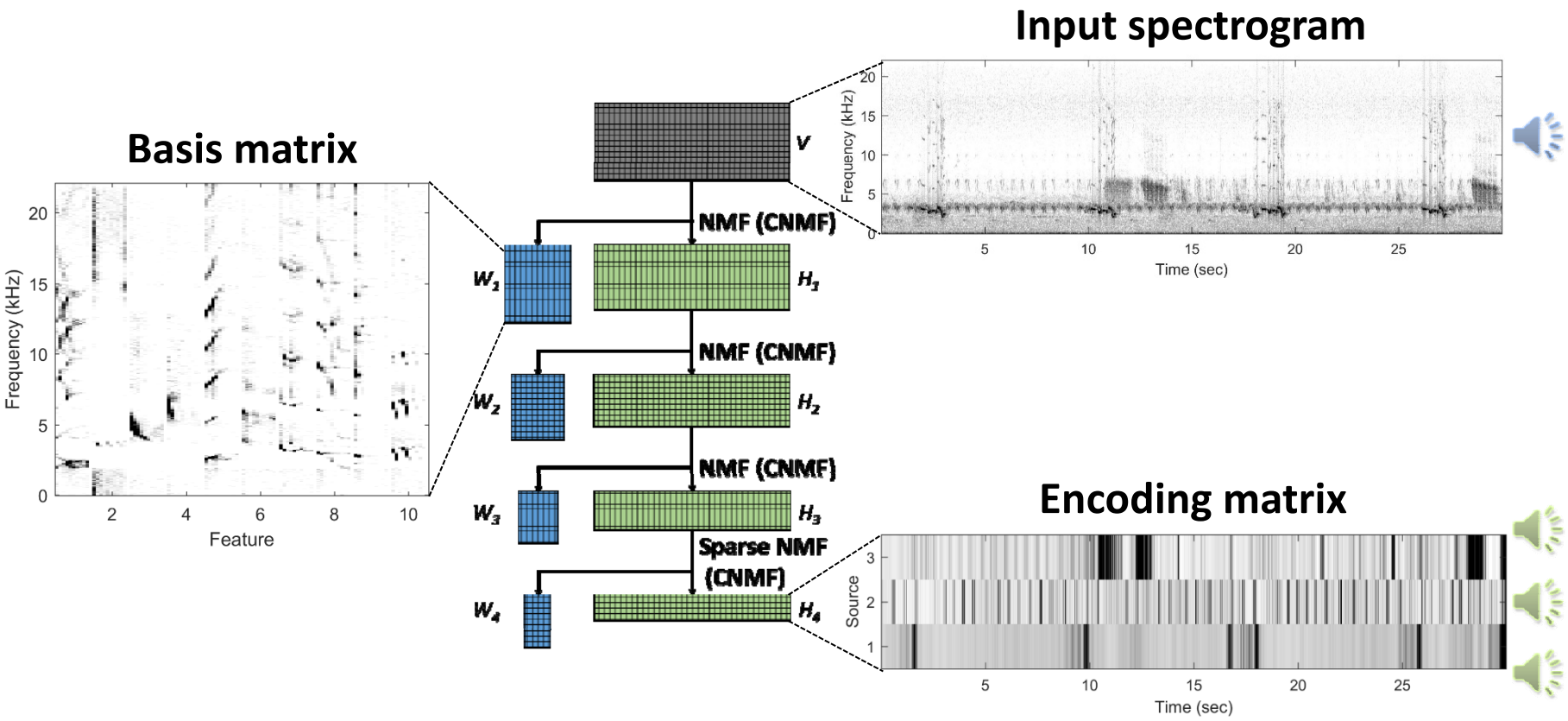
$$H_{a\mu} \leftarrow H_{a\mu} \sum_{i} W_{ia} \frac{V_{i\mu}}{(WH)_{i\mu}}$$

$$H = \underset{H>0}{\operatorname{argmin}}\, d(V, WH) + \lambda \|H\|_1$$

Sparsity constraint

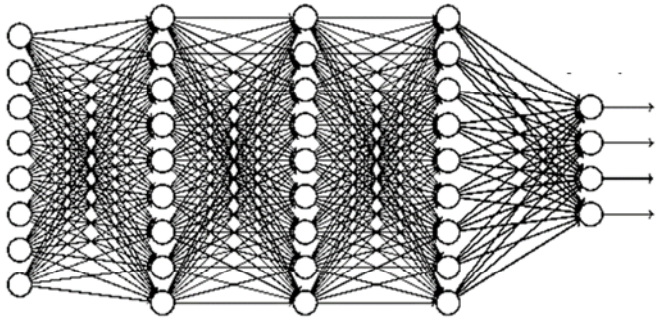
Multiple layers of NMF

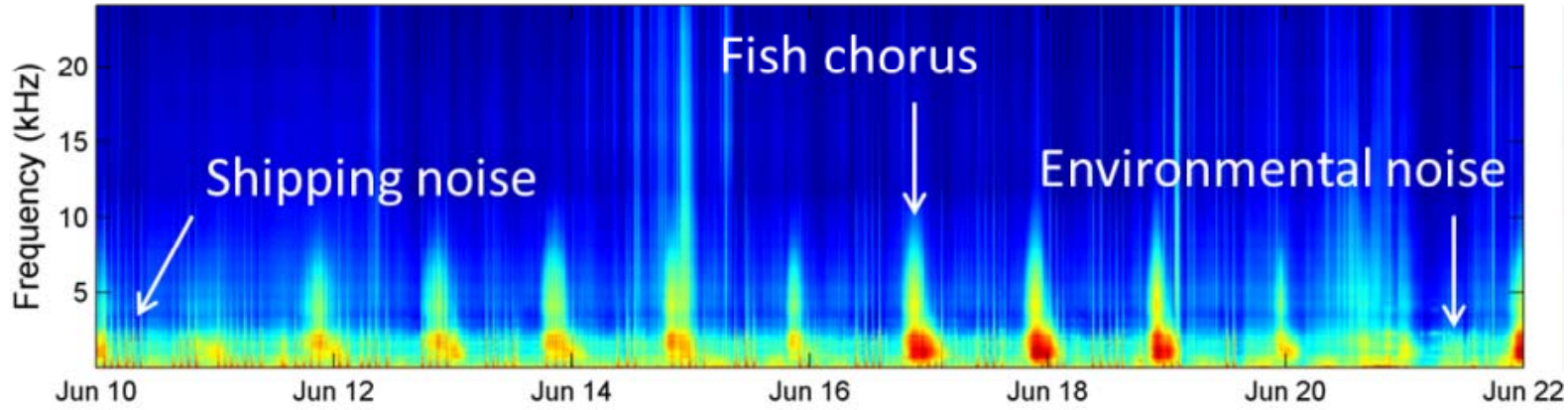
- Estimate basis weights by multiple layers of NMF or CNMF
- Learn the encoding information of *k* sources by the sparse layer



Objectives

- Applying the MLNMF to…
  - Separate soundscape components from long-duration recordings
  - Separate different species of animal vocalizations
  - Search target signals from noisy recordings

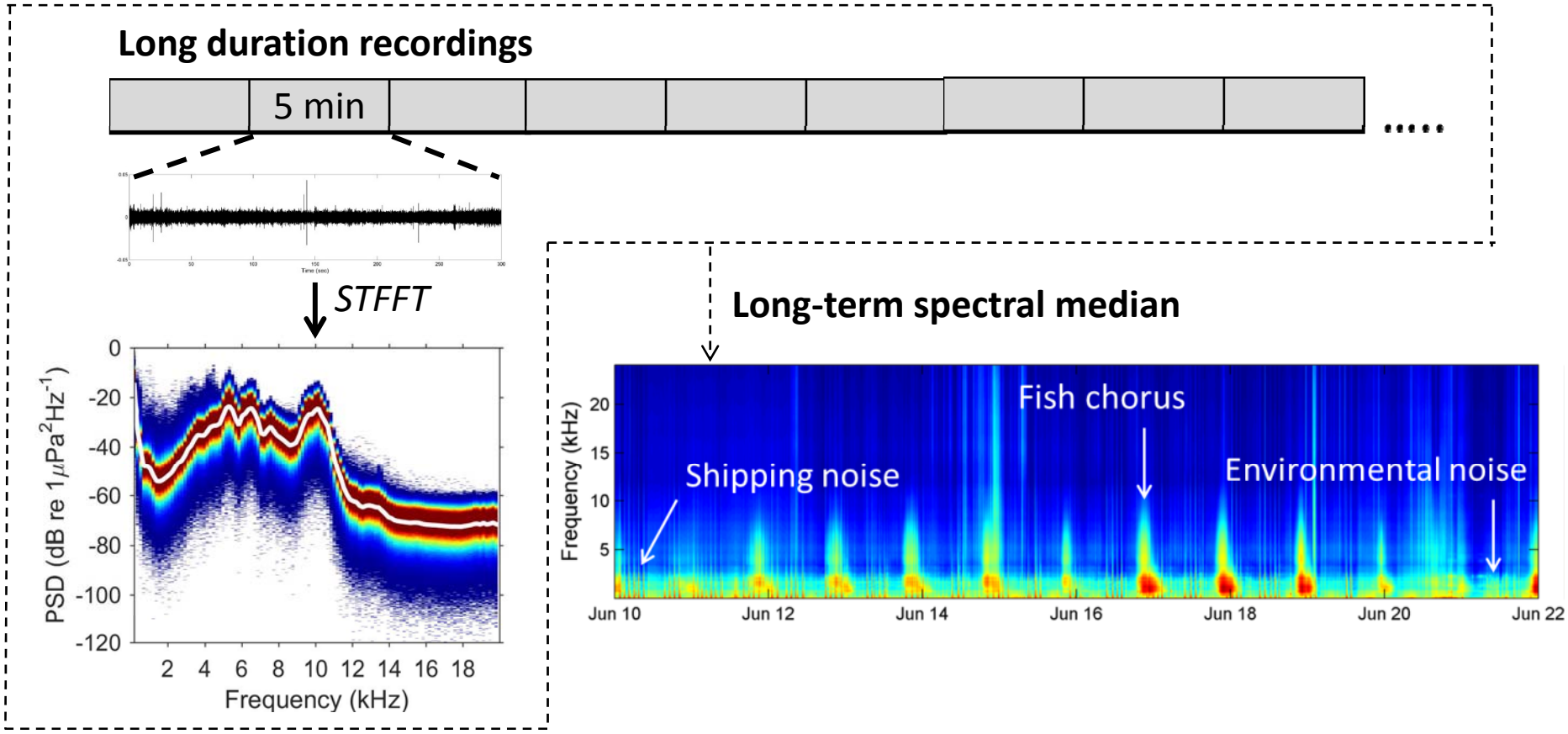




SEPARATION OF SOUNDSCAPE COMPONENTS

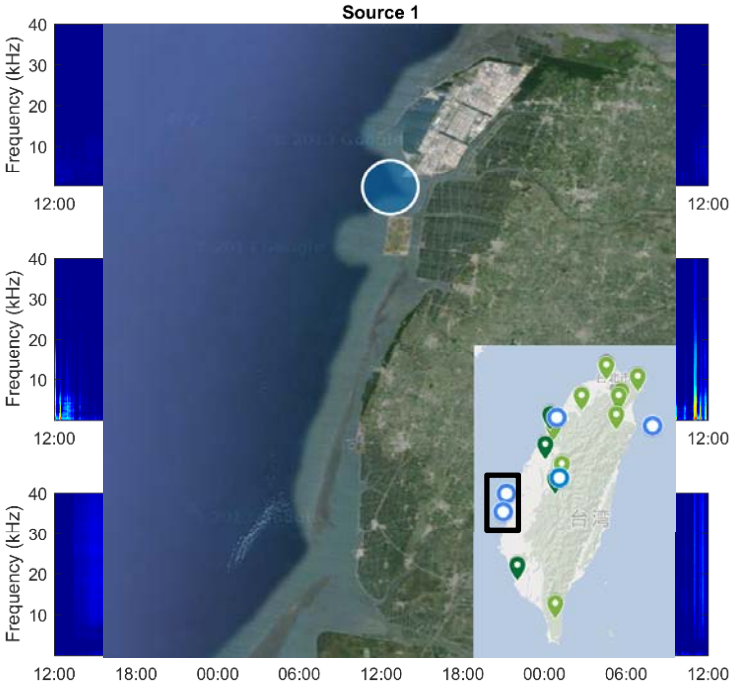
Visualization of long-duration recordings

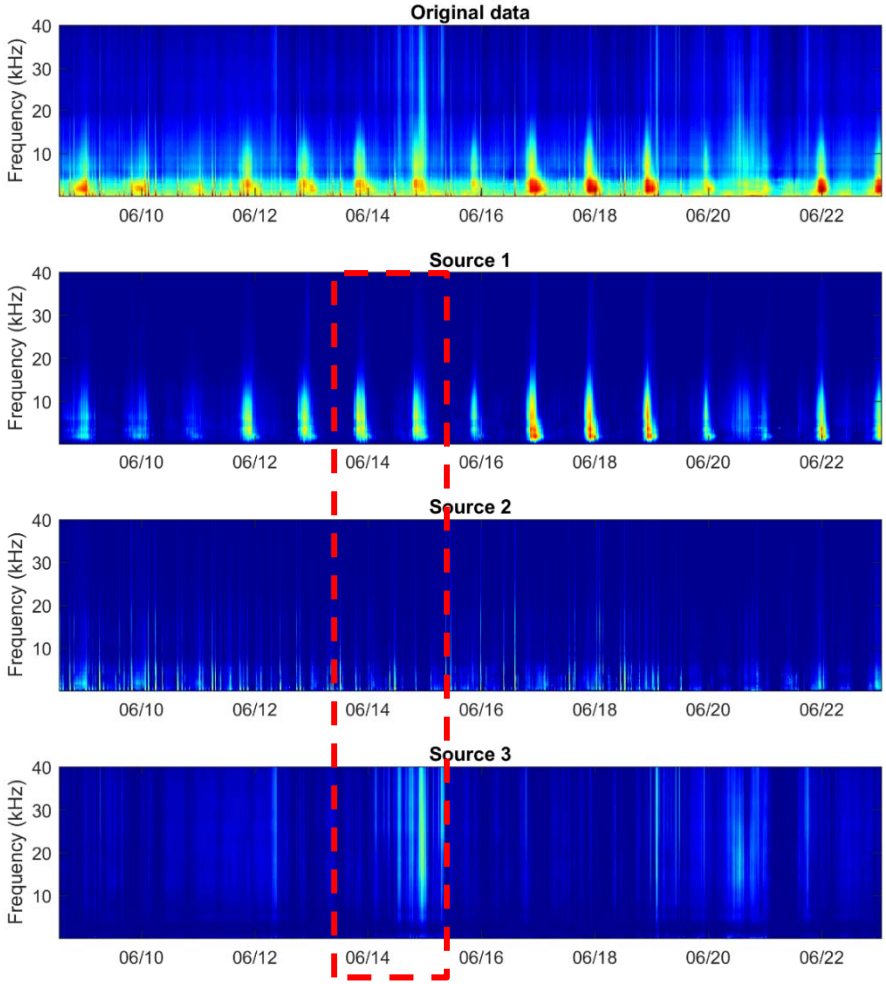
- Visualization of marine soundscape using a long-term spectrogram



BSS of biophony and anthrophony

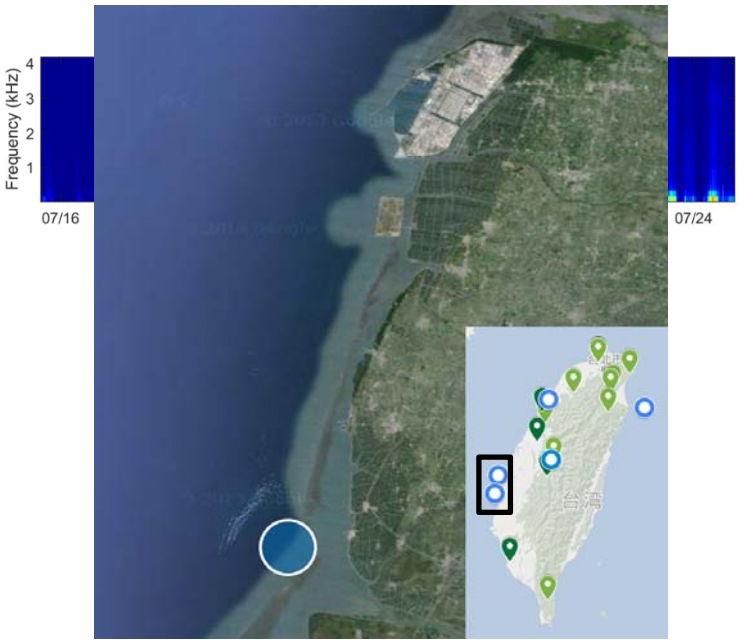
- Recordings collected near by the Mailiao industrial harbor, Yunlin County
  - Fish chorus
  - Shipping noise
  - Snapping shrimp sounds

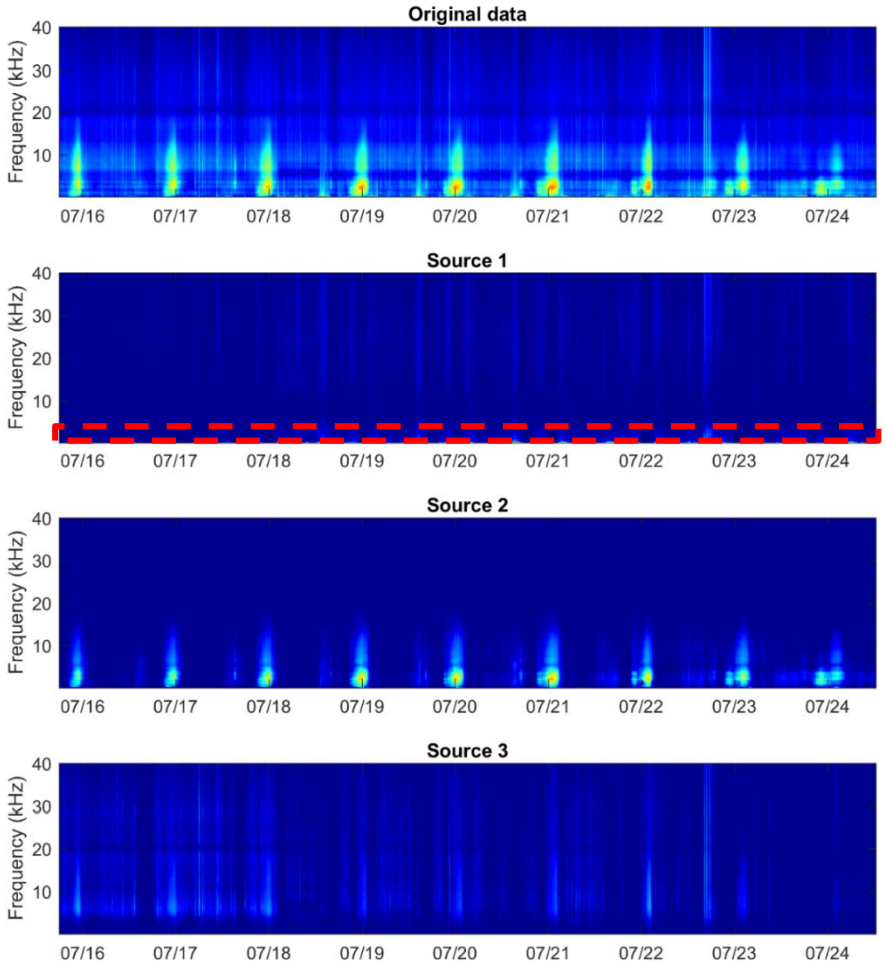


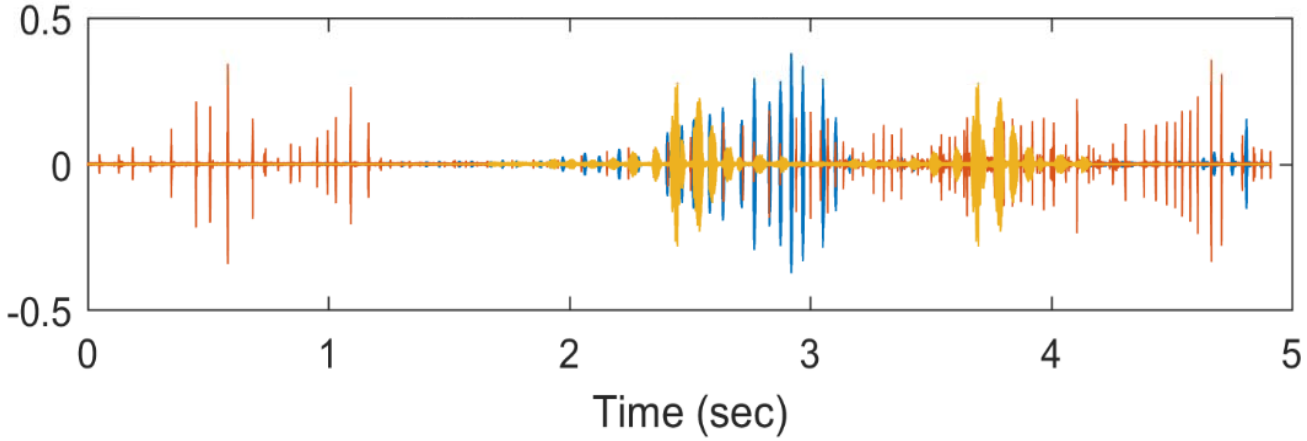


BSS of biophony and geophony

- Recordings collected near by Waisanding sandbar, Yunlin County
  - **Current noise**
  - **Fish chorus**
  - **Snapping shrimp sounds**



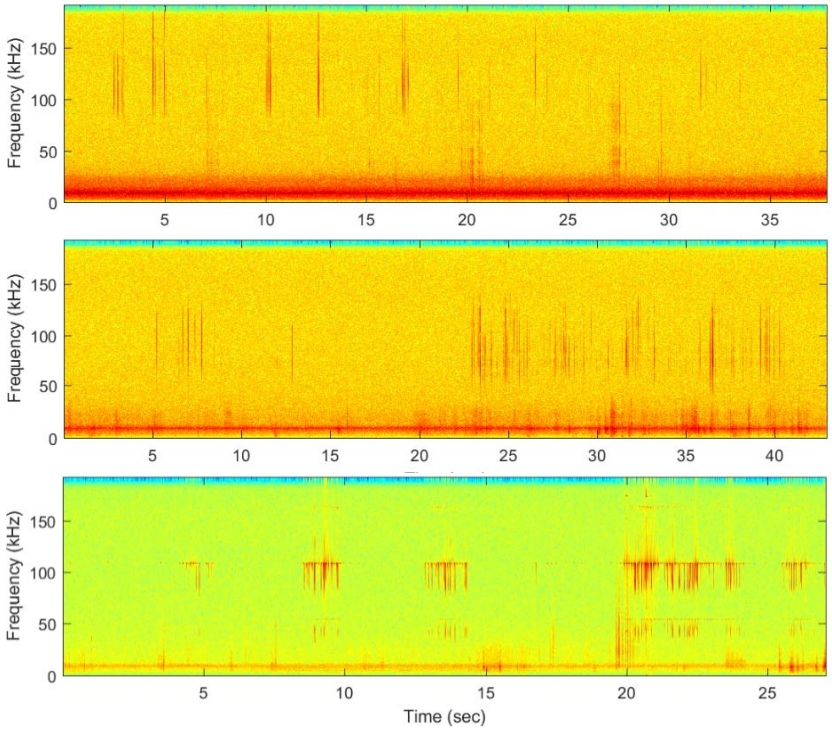




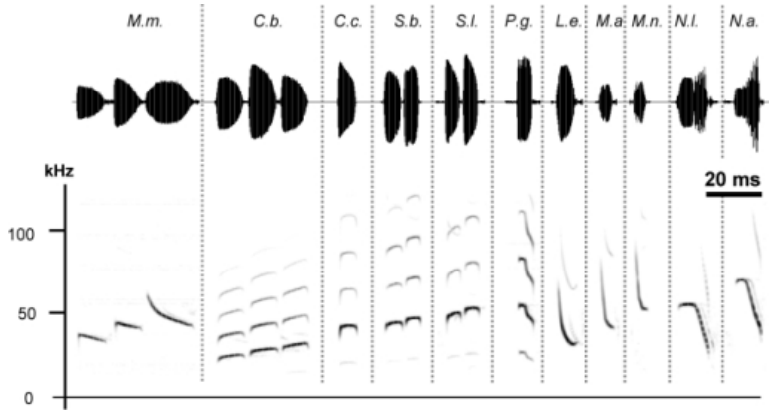
SEPARATION OF SPECIES-SPECIFIC CALLS

Separation of different bat echolocation calls

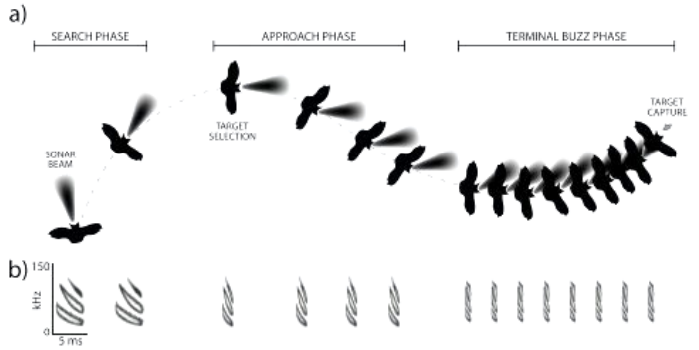
- Reference calls of 13 bat species, that differed in spectral and temporal features



Data collected by Biodiversity Research Center, Academia Sinica, Taiwan



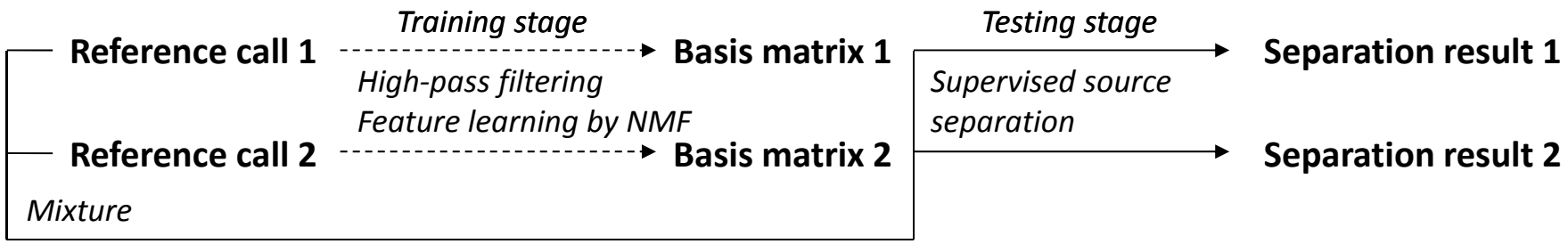
Surlykke and Kloko (2008) PLOS ONE



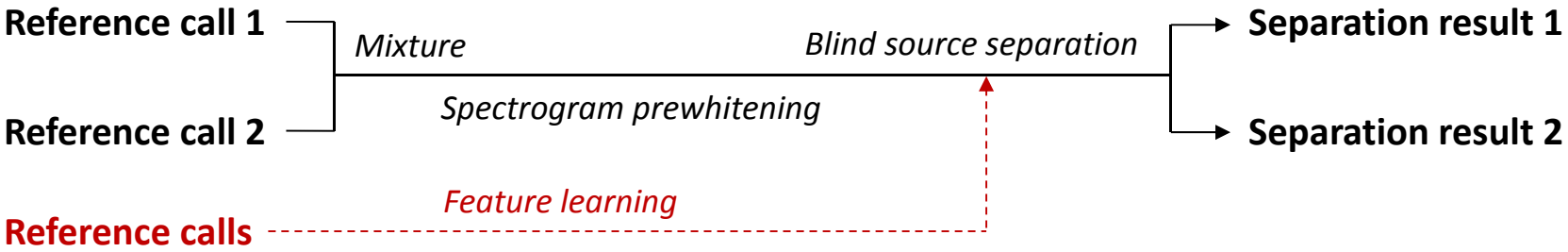
http://tcabasa.org/?page_id=2718

Deep NMF-BSS outperformed NMF-based SS

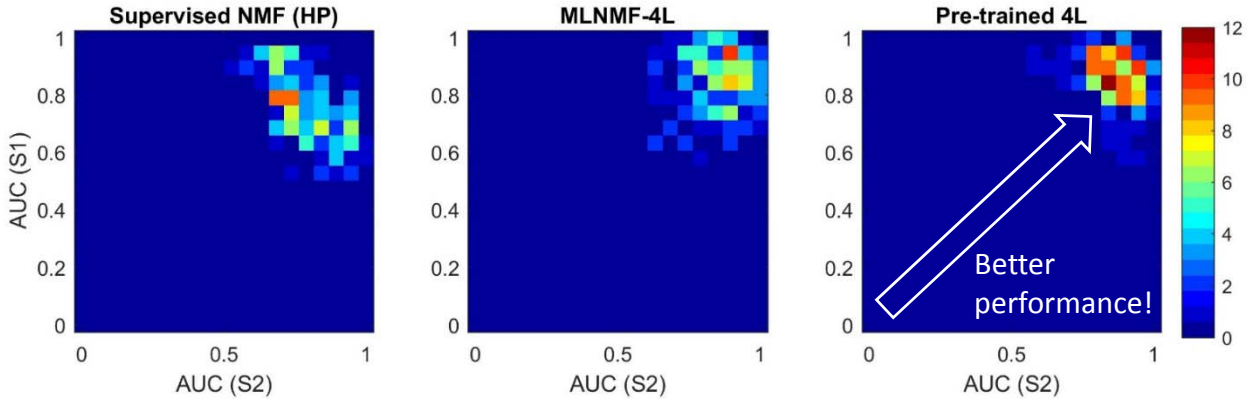
- **Supervised NMF**



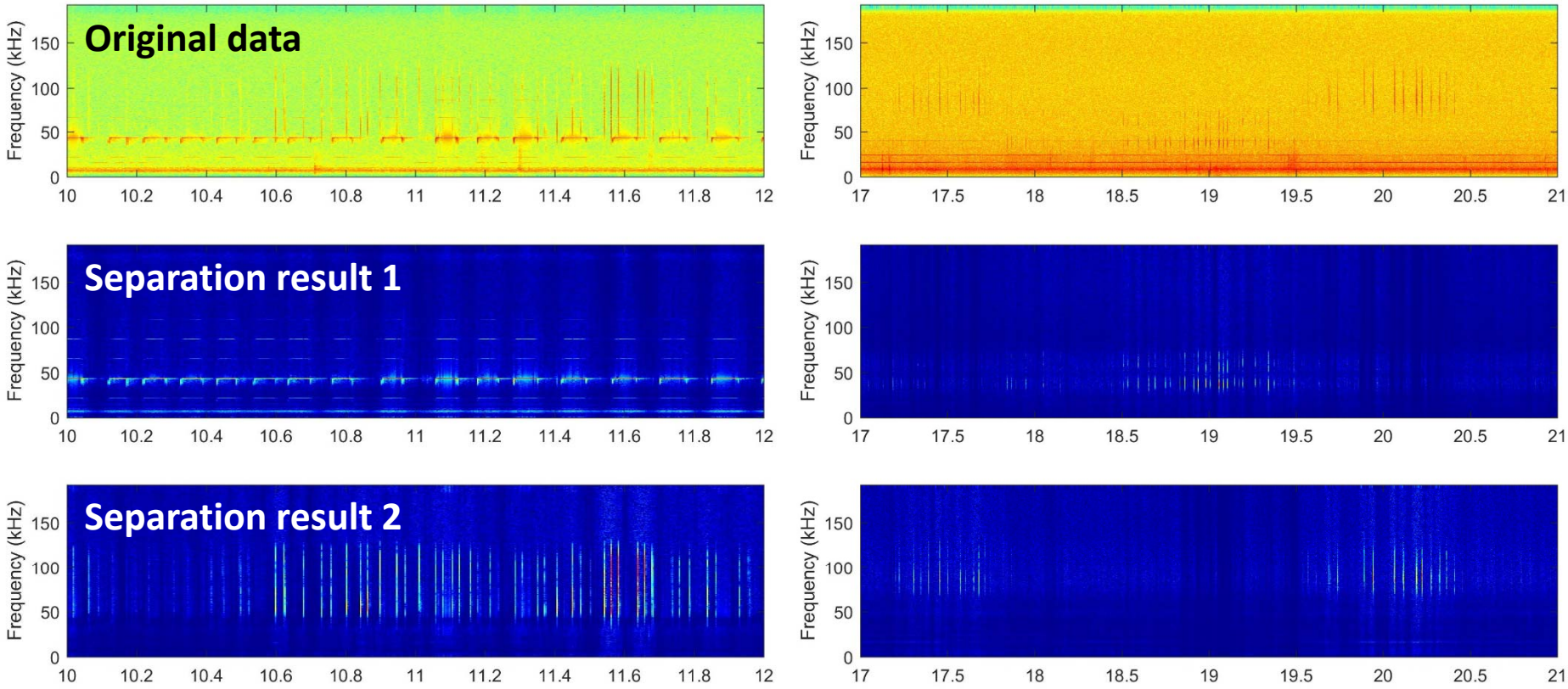
- **MLNMF-4L (pre-trained)**

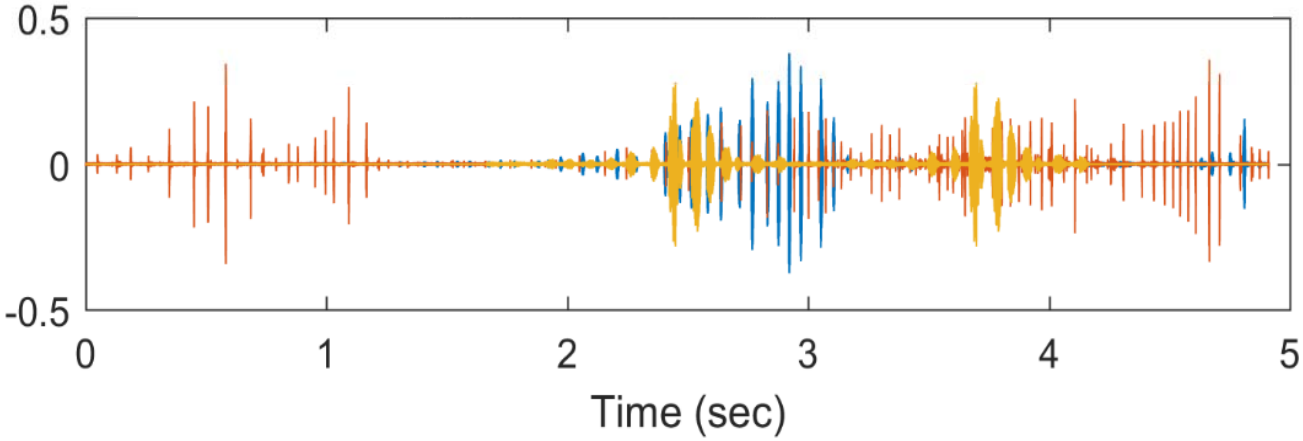


Distribution of AUC scores



Examples of separation result



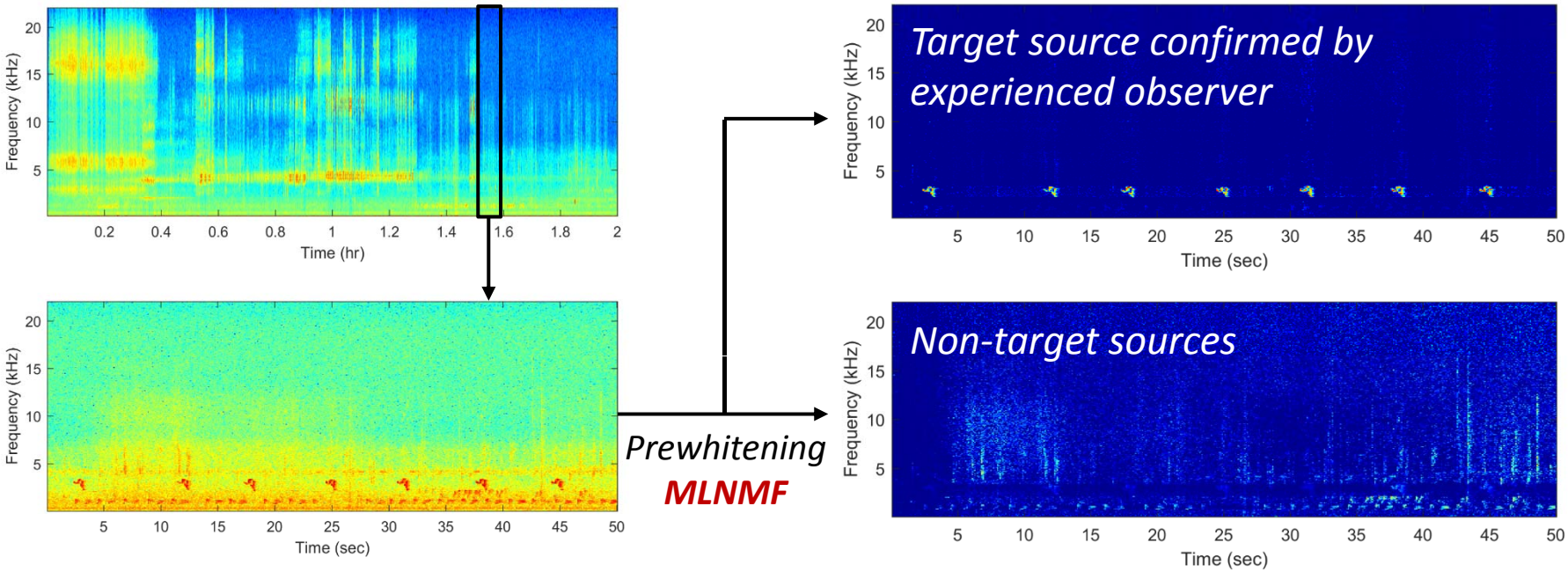


SEARCHING IN A NOISY SOUNDSCAPE

Bird calls in a subtropical forest

- **Dusky fulvetta (*Schoeniparus brunnea*)**
  - Common species < 2000 m elevations

Field recordings from 5 AM – 7 AM



Data collected by National Museum of Natural Science, Taiwan

Performance of “audio search”

Prior knowledge based on the MLNMF model trained by 50 sec

Pre-trained basis matrices

Encoding matrices with random initials

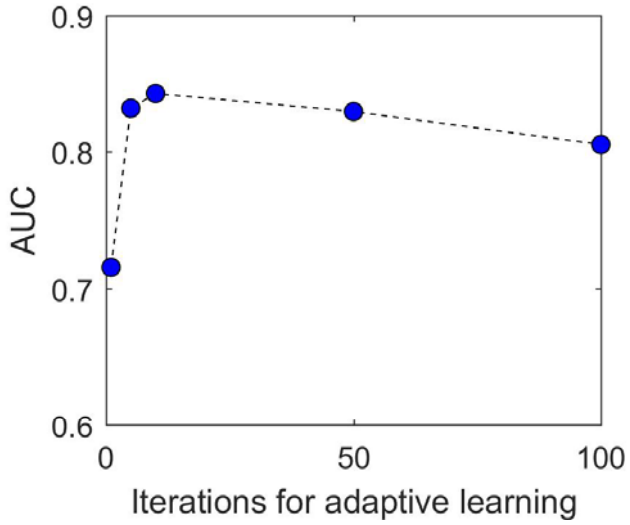
Adaptive learning in new testing data

Adapted basis matrices

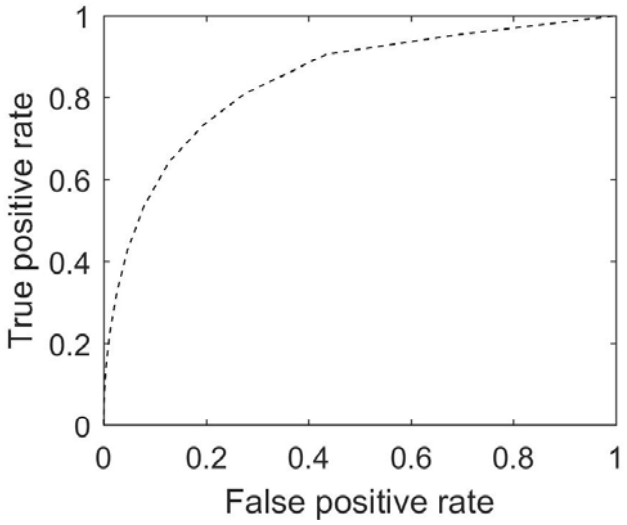
Adapted encoding matrices

Reconstruction of target source

Separation performance of using different iterations

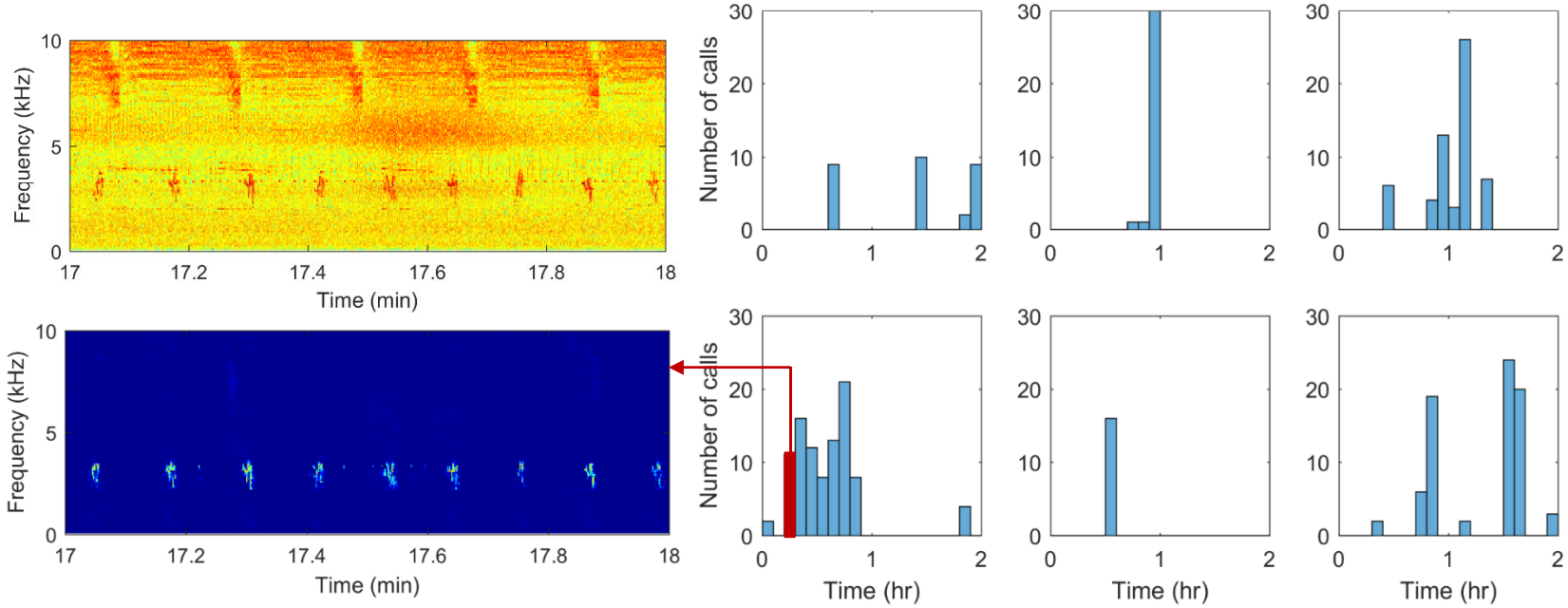


Detection performance of 10 iterations of adaptive learning



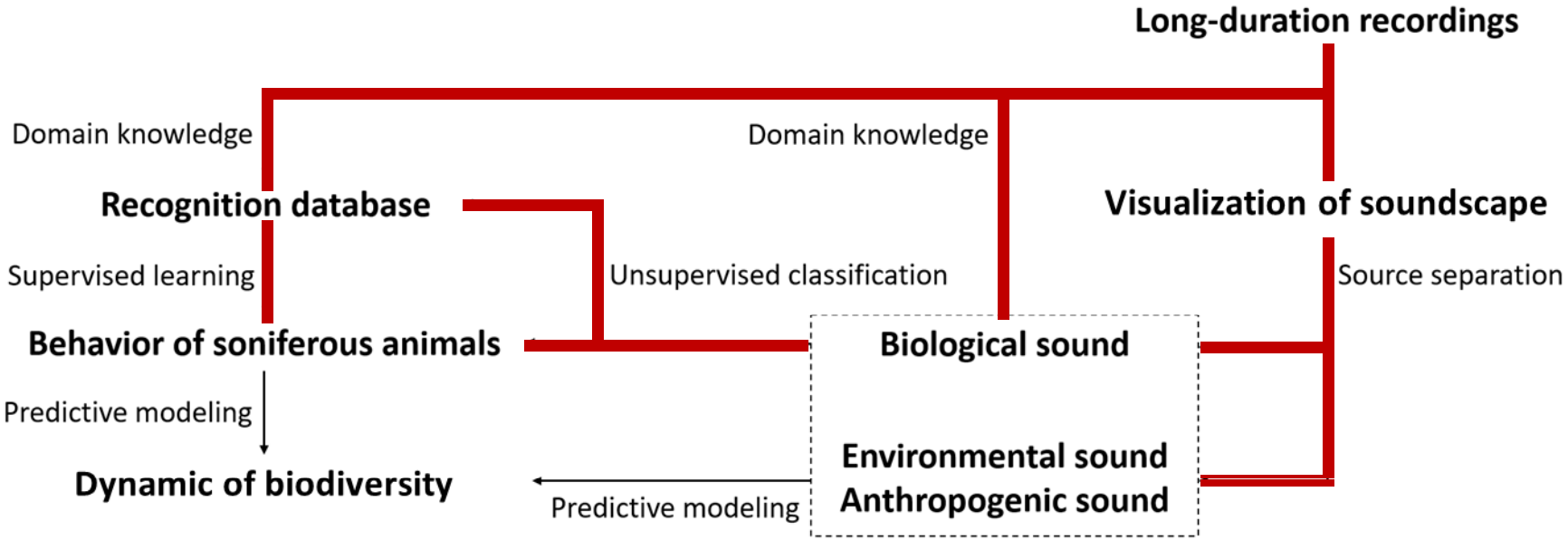
Spatial-temporal distribution of bird calls

- MLNMF model perform well in searching the target bird call by using a small training data (unlabeled!)
- Noise types not encountered in the training data can also be separated effectively



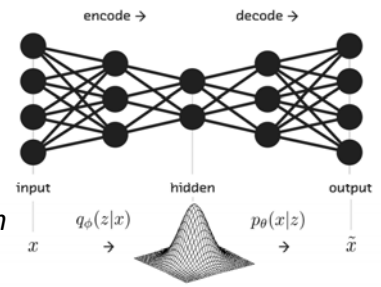
Improving biodiversity monitoring using SIR

- Reduce the difficulty of acoustical analysis for ecologists
  - Effective blind source separation
    - Soundscape components on a long-term spectrogram
    - Animal vocalizations with different spectral and temporal characteristics
  - Efficient supervised separation by using a small training data

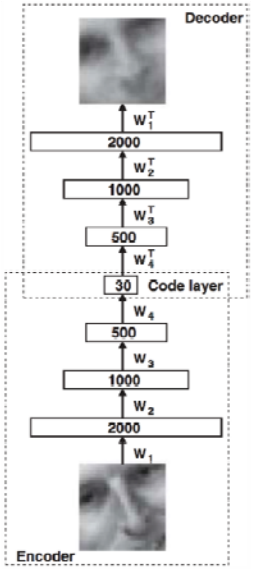


Future integration of deep learning and SIR

- **Unsupervised learning (Deep AE, Variational AE)**
  - Improve NMF-based blind source separation
  - Facilitate the collection of labeled data with minimum manpower?

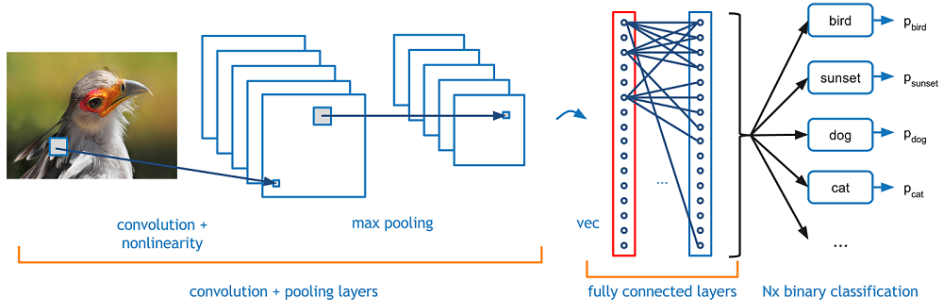


blog.fastforwardlabs.com



Hinton & Salakhutdinov (2006) Science

- **Supervised learning (CNN, RNN…)**
  - Identification of species by learning from a large amount of training data
  - Identification of behavior or individual?

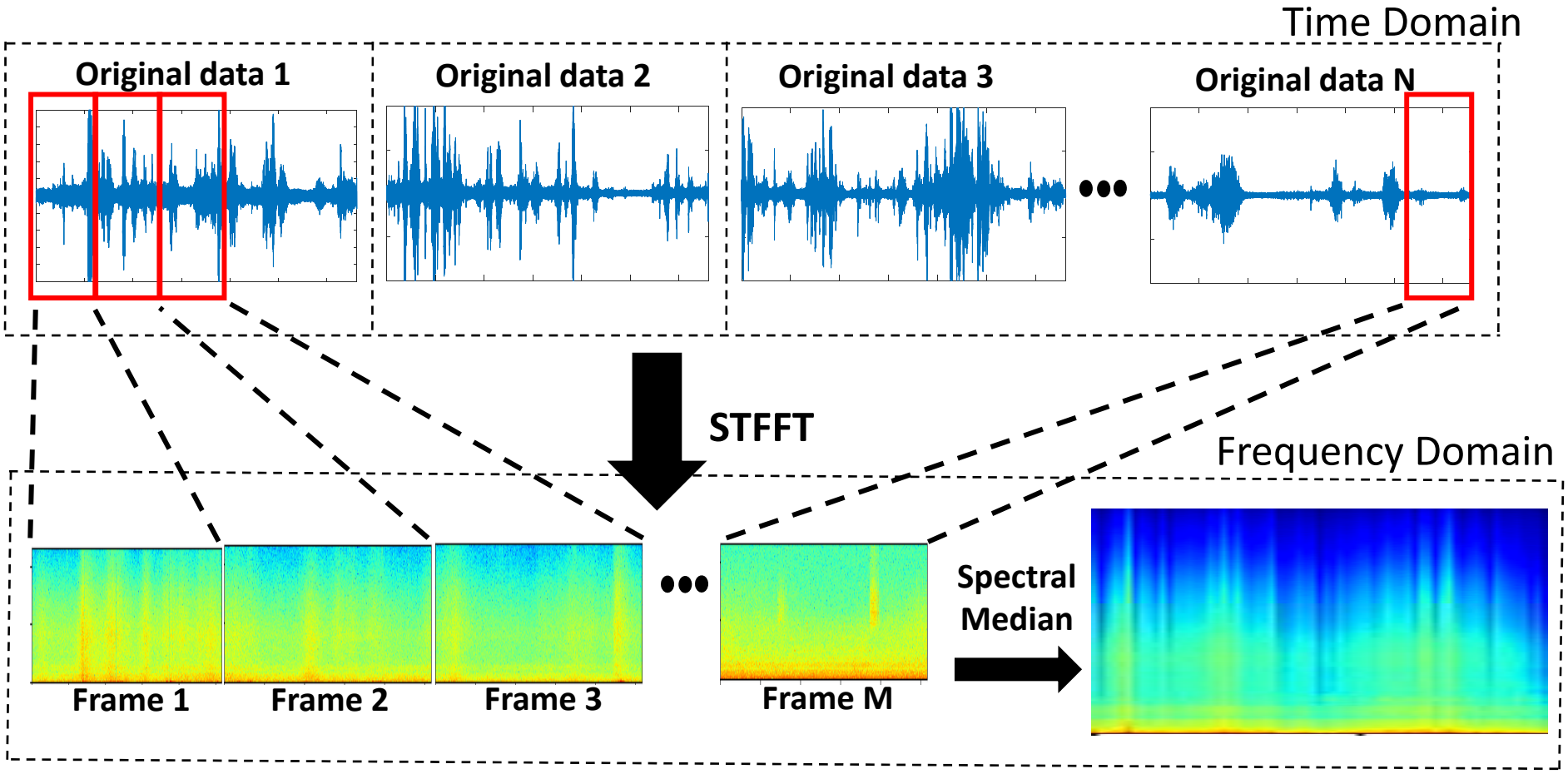


adeshpande3.github.io

Welcome for collaboration!

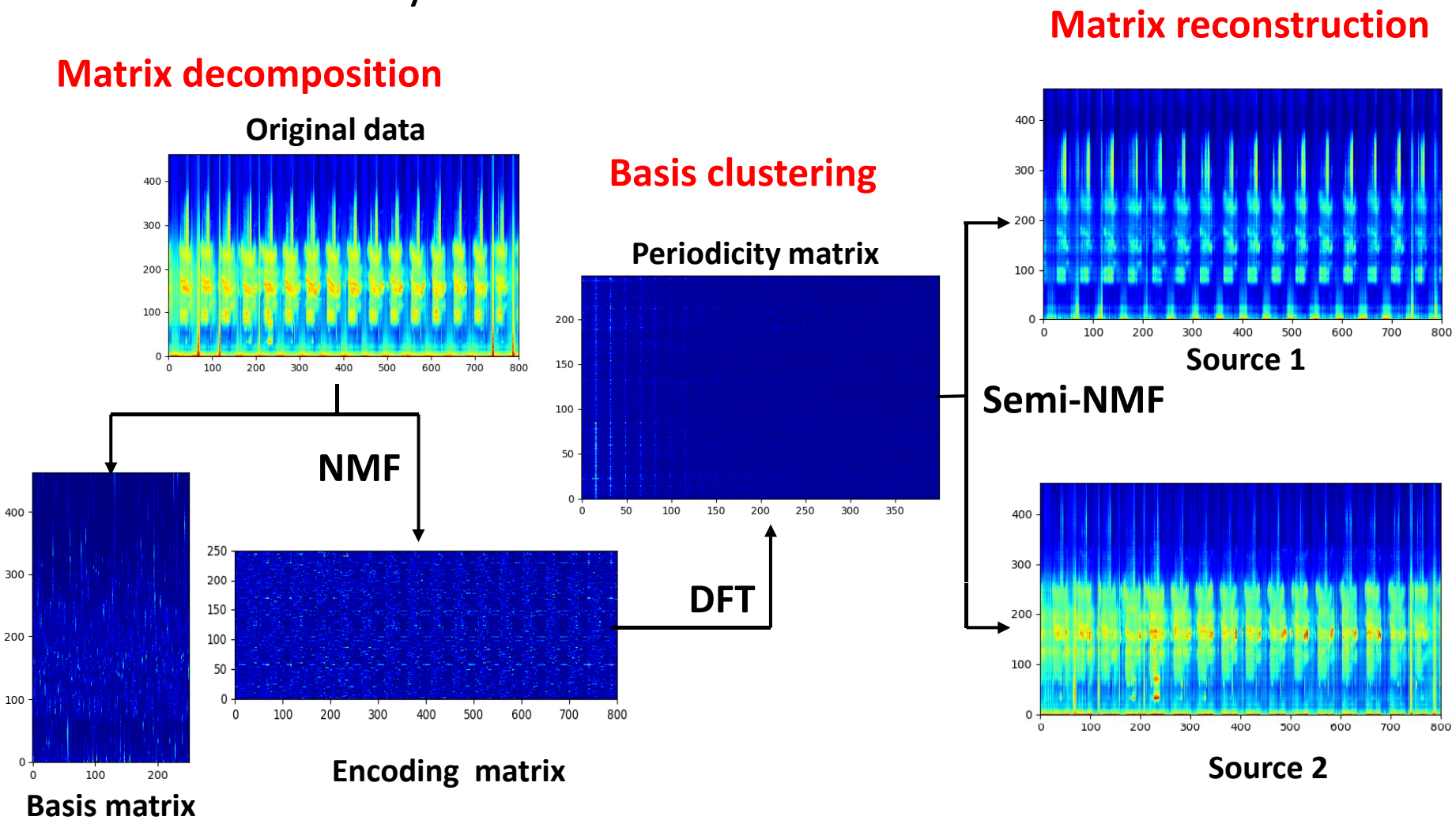
Visualization of long-term spectgoram

- Reduce the redundant information by measuring the median power spectrum



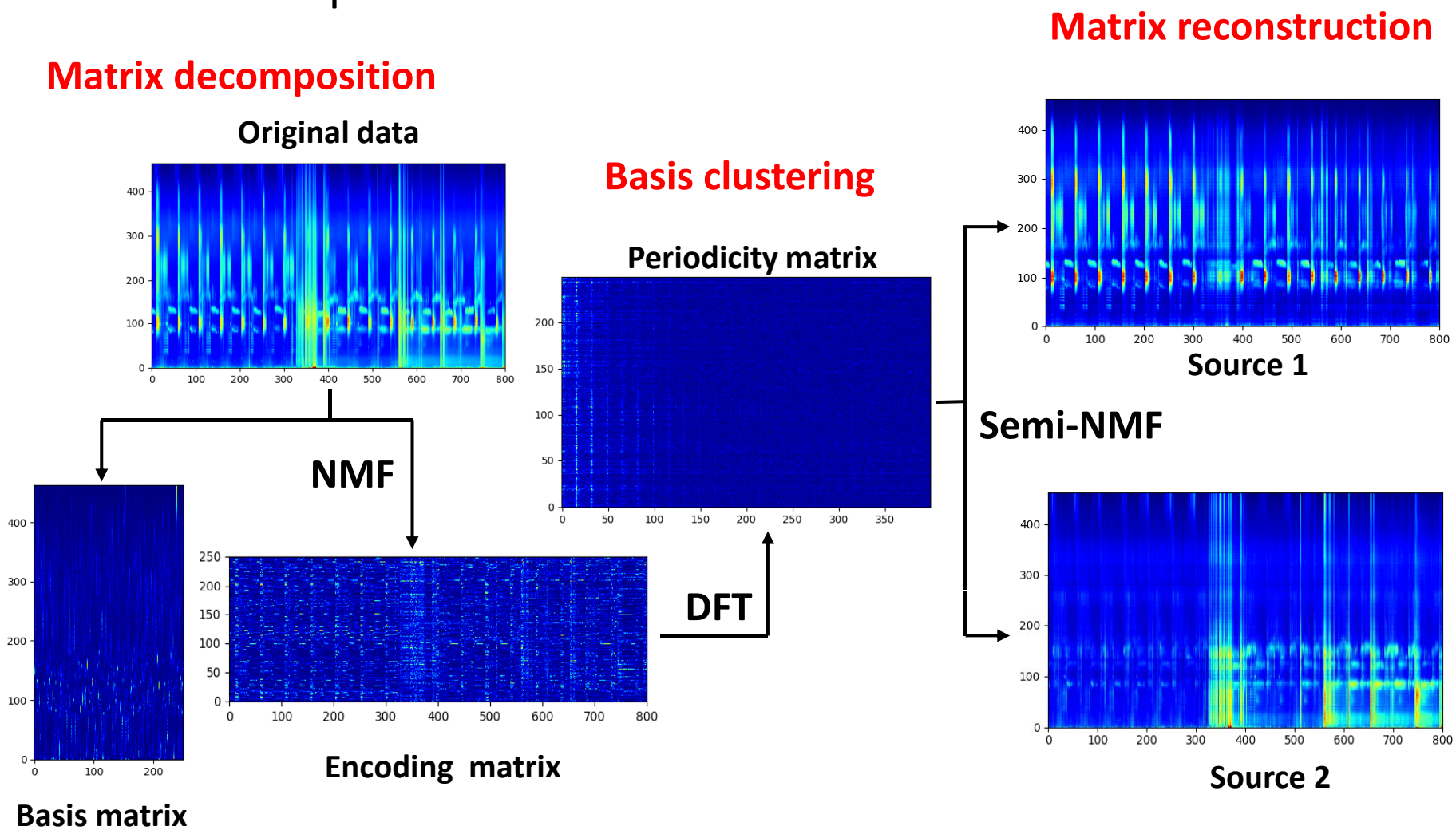
Visualization of PCNMF

- Terrestrial in noisy



Visualization of PCNMF

- Terrestrial in quiet



- Codes are available: goo.gl/4famVt

References

- T.-H. Lin, T. Akamatsu, and Y, Tsao, "Comparison of passive acoustic soniferous fish monitoring with supervised and unsupervised approaches," to appear in?Journal of the Acoustical Society of America (JASA).
- T.-H. Lin, S.-H. Fang, and Y, Tsao, "Improving Biodiversity Assessment via Unsupervised Separation of Biological Sounds from Long-duration Recordings," Scientific Reports, volume7, number4547, pages1, July 2017.
- T.-H. Lin, Y.-H. Wang, S.-S. Lu, H.-W. Yen, and Y. Tsao, "Computing Biodiversity Change via a Soundscape Monitoring Network," PNC 2017 Annual Conference and Joint Meetings, November 2017.

Acknowledgement

- Ministry of Science and Technology
  - Independent postdoctoral research fellowship
- Academia Sinica Data Science Seed Grant Program