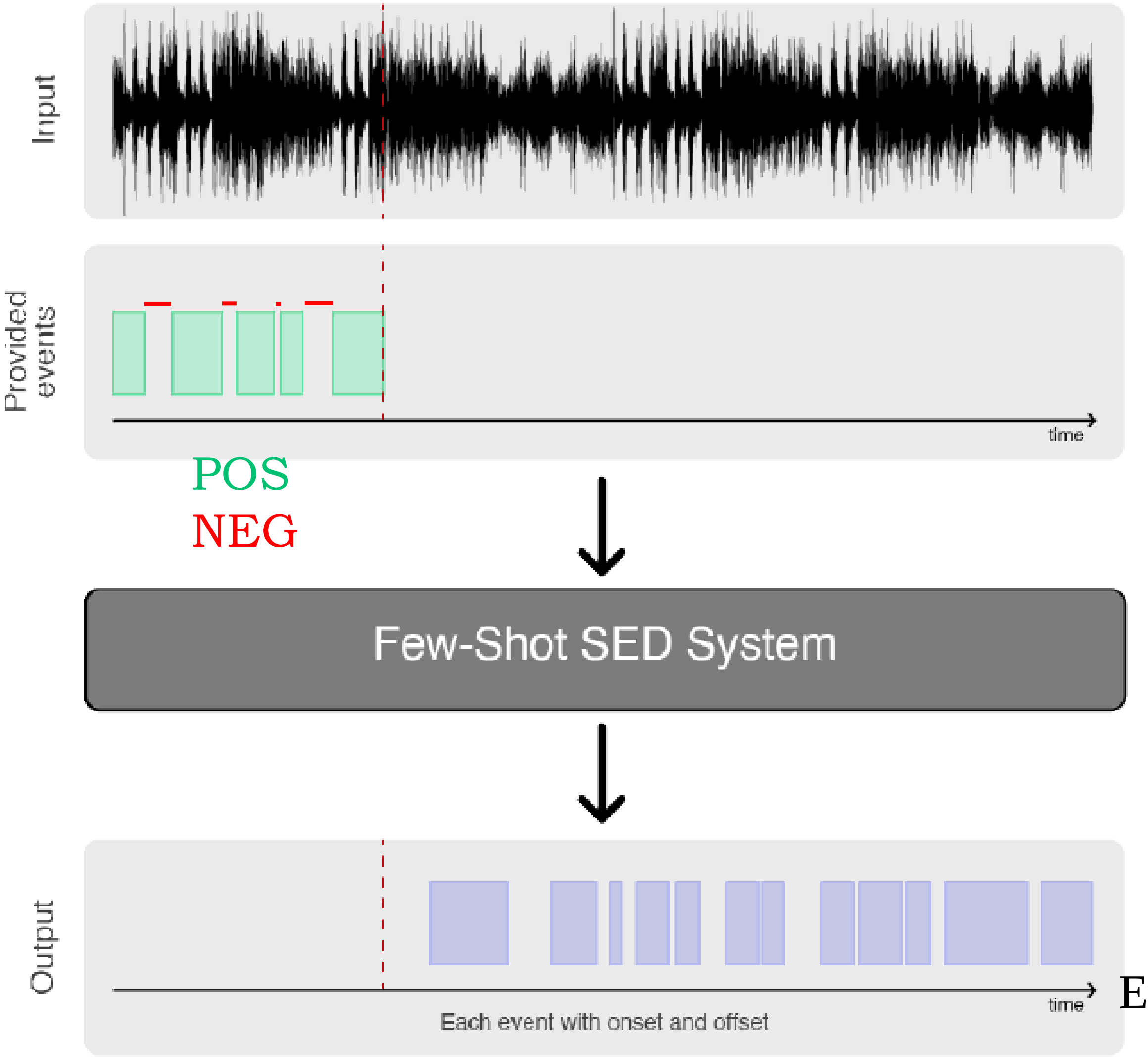
Cosine Similarity-based **Few-shot Bioacoustic Event Detection** with Automatic Frequency Range Identification in Mel-Spectrograms

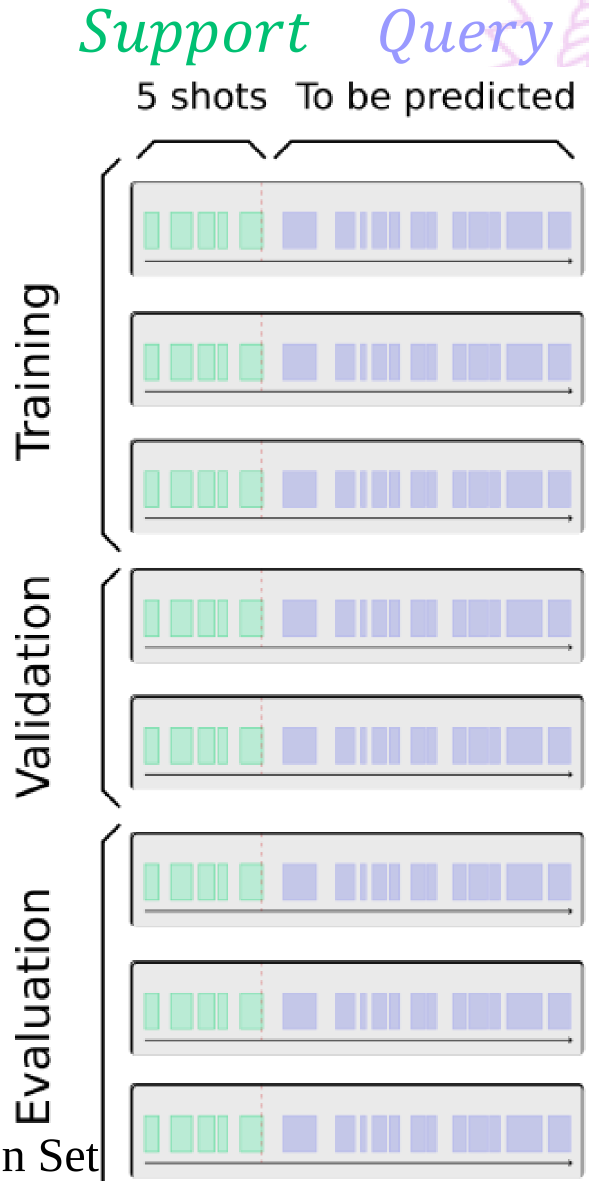
Student: Sheng-Lun Kao

Adviser: Dr. Yi-Wen Liu

Date: 2025.06.26

DCASE 2024 Few-shot Bioacoustic Event Detection: Task description





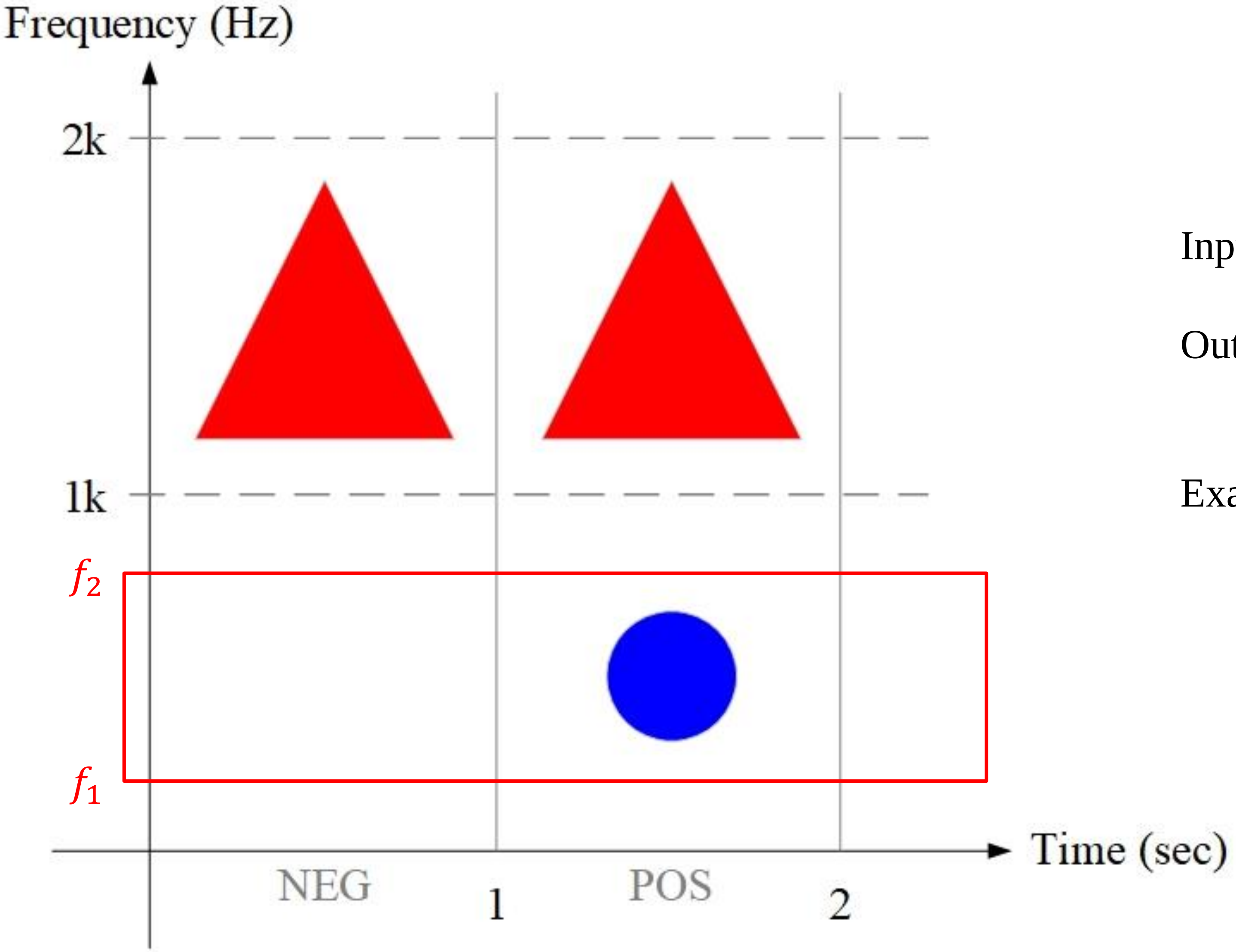
Evaluation Set

only the first 5 annotations are provided for each file.

Goal

- In response to the Few-shot Bioacoustic Event Detection challenge, we have developed a detection system comprising three key components.
  - First, we devised an algorithm called automatic frequency range identification (AFRI), which automatically identifies the frequency range of positive (POS) events within the mel-spectrogram.
  - Secondly, the cosine similarity between POS and negative (NEG) events is computed across the entire audio file.
  - Thirdly, predictions of POS events are made based on the results of cosine similarity.

What is AFRI ?



Input: annotation + Mel-spectrogram

Output: $[f_1, f_2]$

Example of an annotation file for audio_val.wav:

Audiofilename,Starttime,Endtime,Q
audio_val.wav,1.3,1.6,POS
.
.

Why AFRI ?

- Identifying smaller objects in images poses significant challenges [14-18]. Smaller objects often occupy fewer pixels in an image, making them difficult to detect. They are easily overlapped by other noise, and dimensionality reduction can result in the loss of critical details.
- In the DCASE 2024 challenge's evaluation set, which includes 66 audio files, POS events in 13 files are akin to small objects in the Mel-spectrogram.

G. Cheng *et al.*, “Towards large-scale small object detection: Survey and benchmarks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Y. Cao *et al.*, “VisDrone-DET2021: The vision meets drone object detection challenge results,” in *Proceedings of the IEEE/CVF International conference on computer vision*, 2021, pp. 2847–2854.

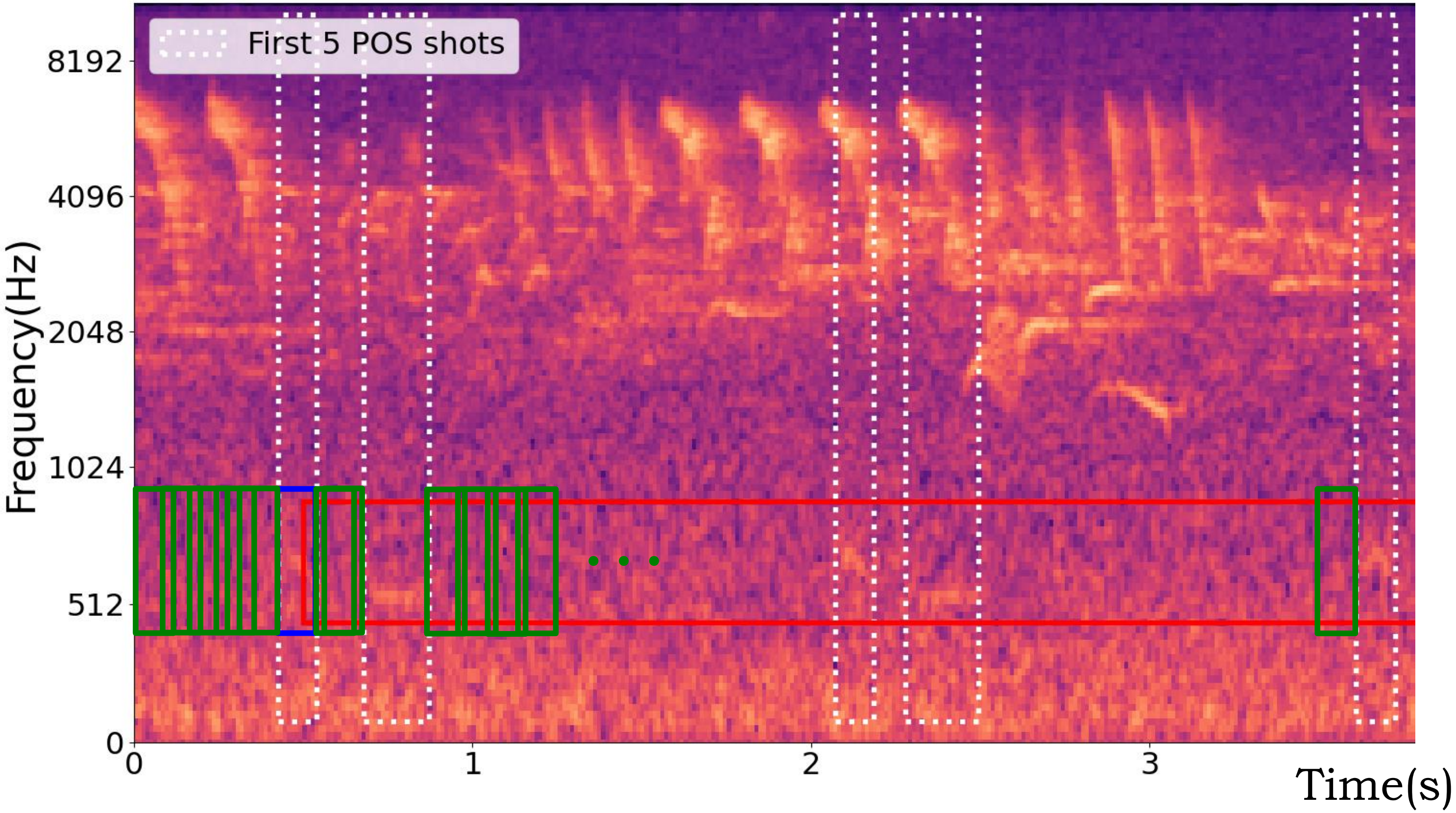
C. Deng, M. Wang, L. Liu, Y. Liu, and Y. Jiang, “Extended feature pyramid network for small object detection,” *IEEE Transactions on Multimedia*, vol. 24, pp. 1968–1979, 2021.

J.-S. Lim, M. Astrid, H.-J. Yoon, and S.-I. Lee, “Small object detection using context and attention,” in *2021 international Conference on Artificial intelligence in information and Communication (ICAIIC)*, 2021, pp. 181–186.

G. Chen *et al.*, “A survey of the four pillars for small object detection: Multiscale representation, contextual information, super-resolution, and region proposal,” *IEEE Transactions on systems, man, and cybernetics: systems*, vol. 52, no. 2, pp. 936–953, 2020.

How AFRI ?

POS – **NEG**(the most similar to POS)



$$\mathrm{cossim}(a_1,b_1) = \beta\big(\mathrm{TMD}(a_1),\mathrm{TMD}(b_1)\big) + \beta\big(\mathrm{FMD}(a_1),\mathrm{FMD}(b_1)\big)$$

cosine similarity function β

time marginal distribution (TMD)

$$\mathrm{TMD}(t;z) = \frac{1}{N_f}\sum_{f=1}^{N_f} S(t,f;z)$$

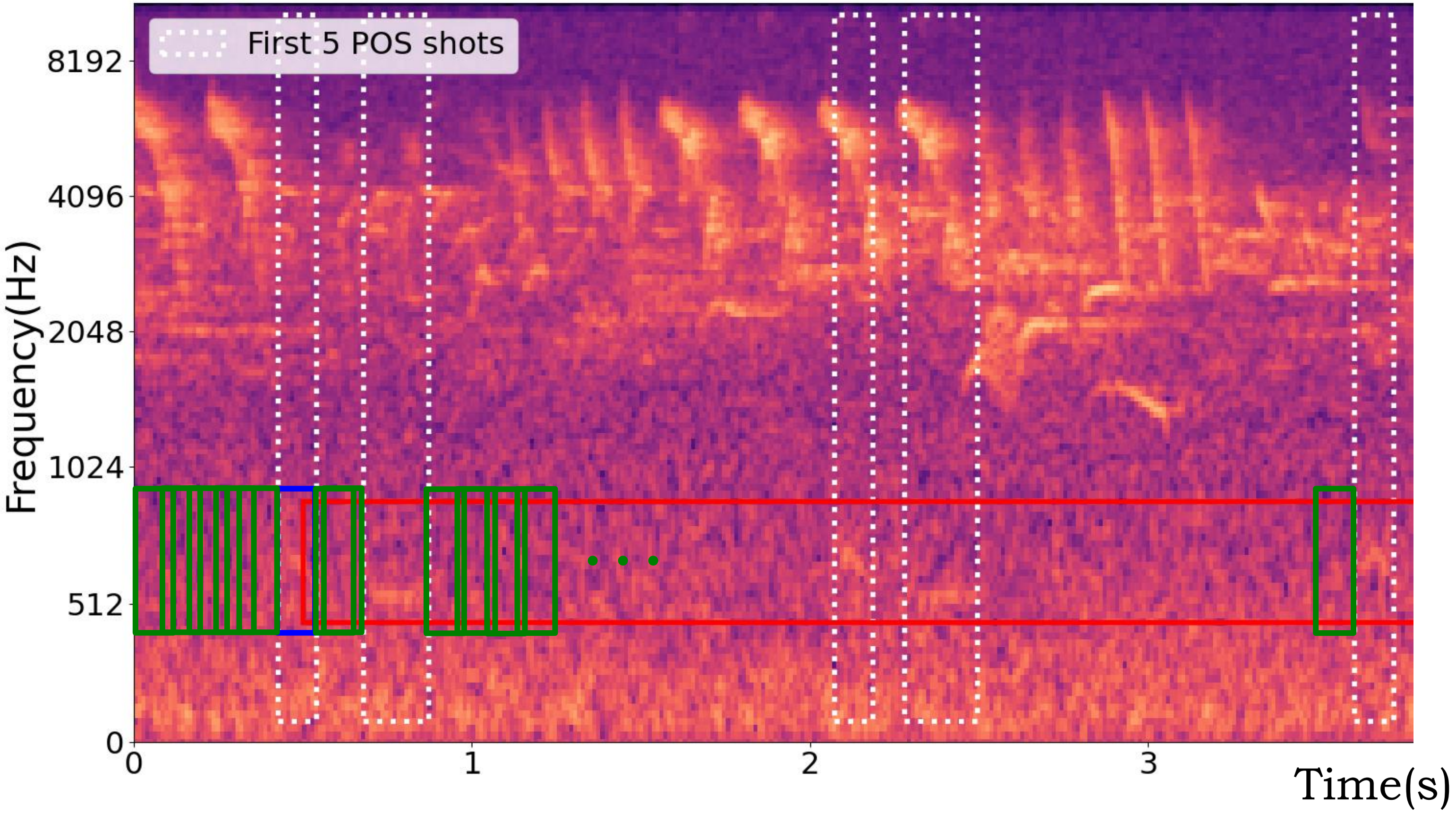
frequency marginal distribution (FMD)

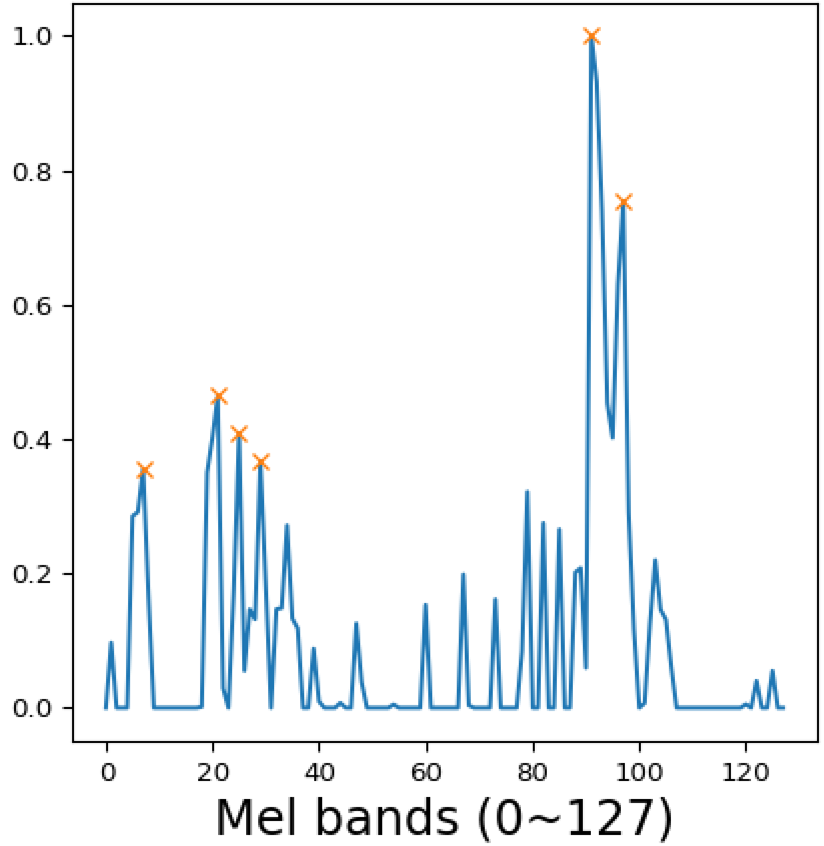
$$\mathrm{FMD}(f;z) = \frac{1}{N_t}\sum_{t=1}^{N_t} S(t,f;z)$$

loc of Max $[\mathrm{cossim}(a_1,b_1)\ \ \mathrm{cossim}(a_1,b_2)\ \ \mathrm{cossim}(a_1,b_3)\ \ \ldots\ \ \mathrm{cossim}(a_1,b_n)]$

How AFRI ?

FMD(POS) – FMD(NEG)(the most similar to POS)

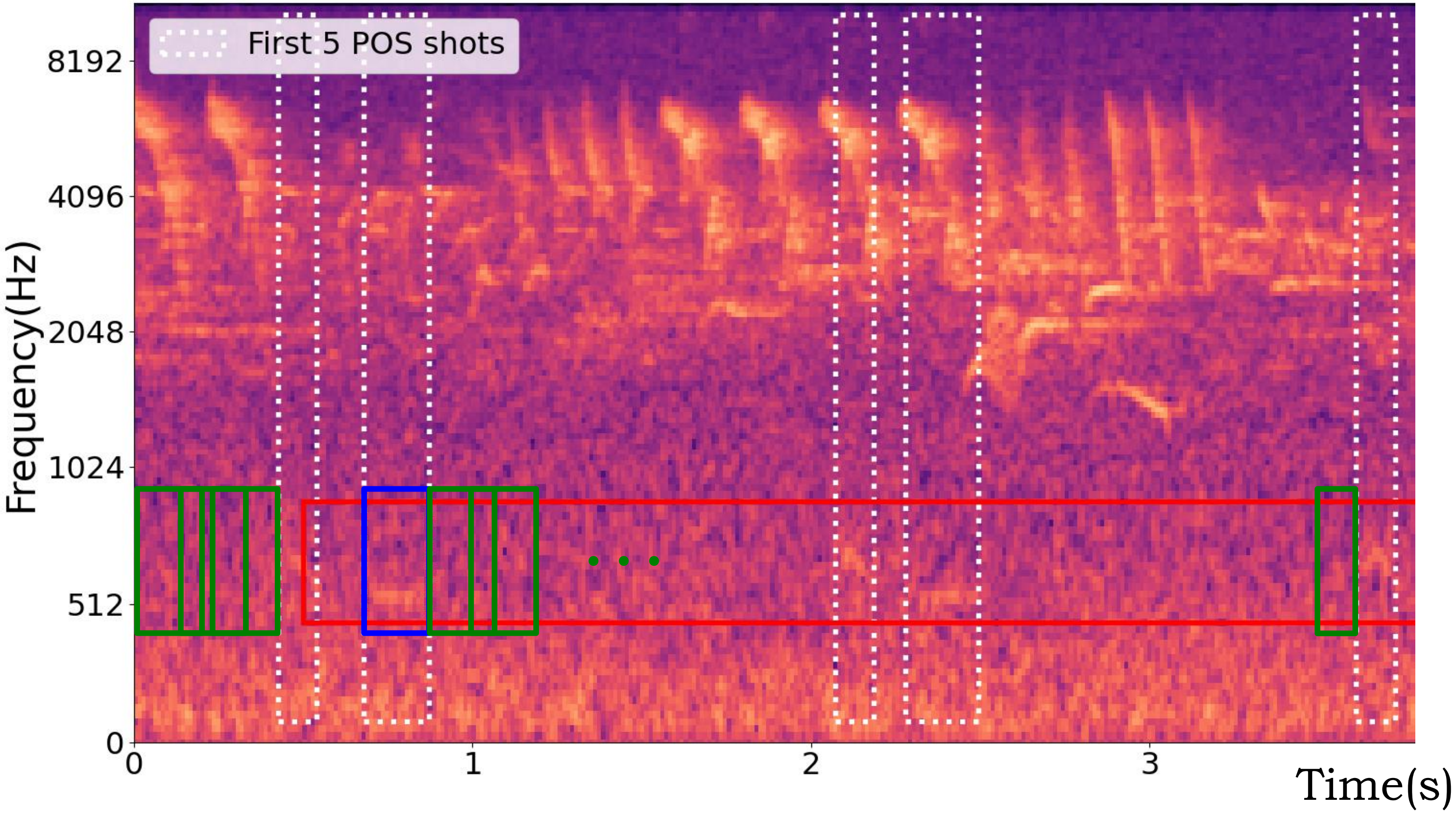


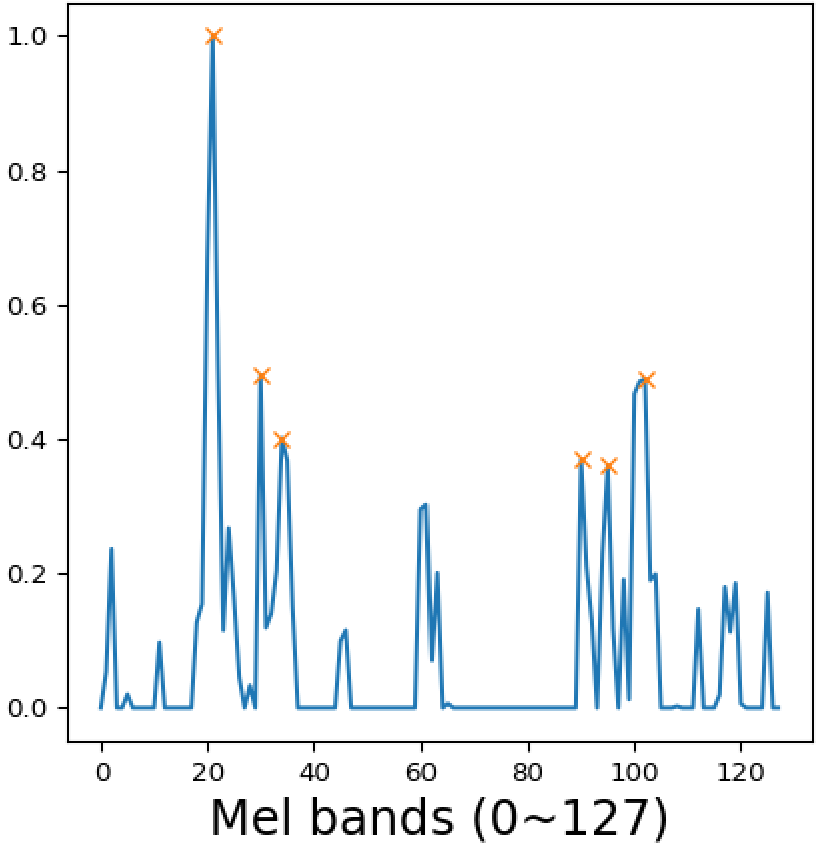


loc of Max [cossim(a_1,b_1) cossim(a_1,b_2) cossim(a_1,b_3) … cossim(a_1,b_n)]

國立清華大學
NATIONAL TSING HUA UNIVERSITY

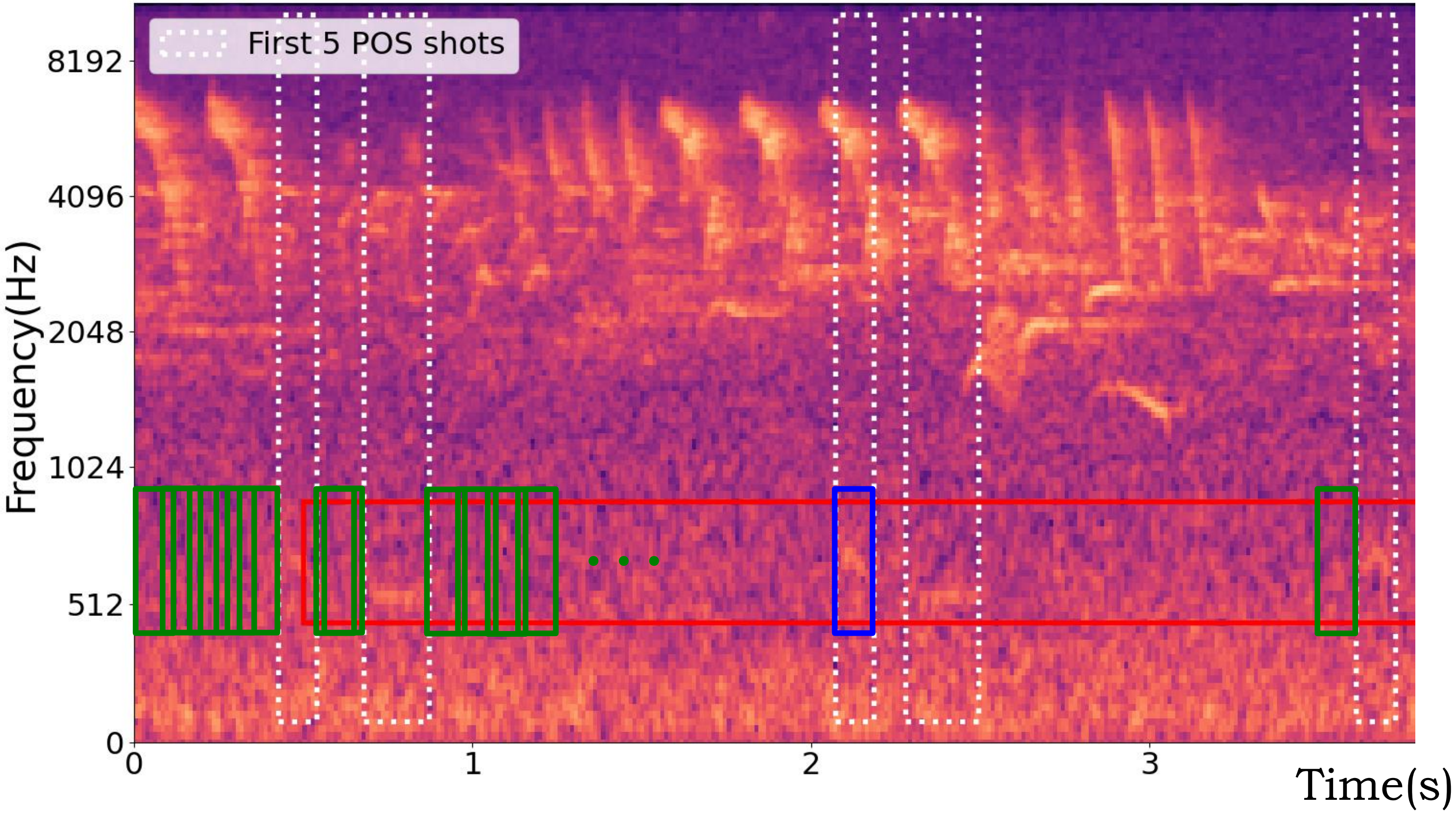
FMD(POS) – FMD(NEG)(the most similar to POS)

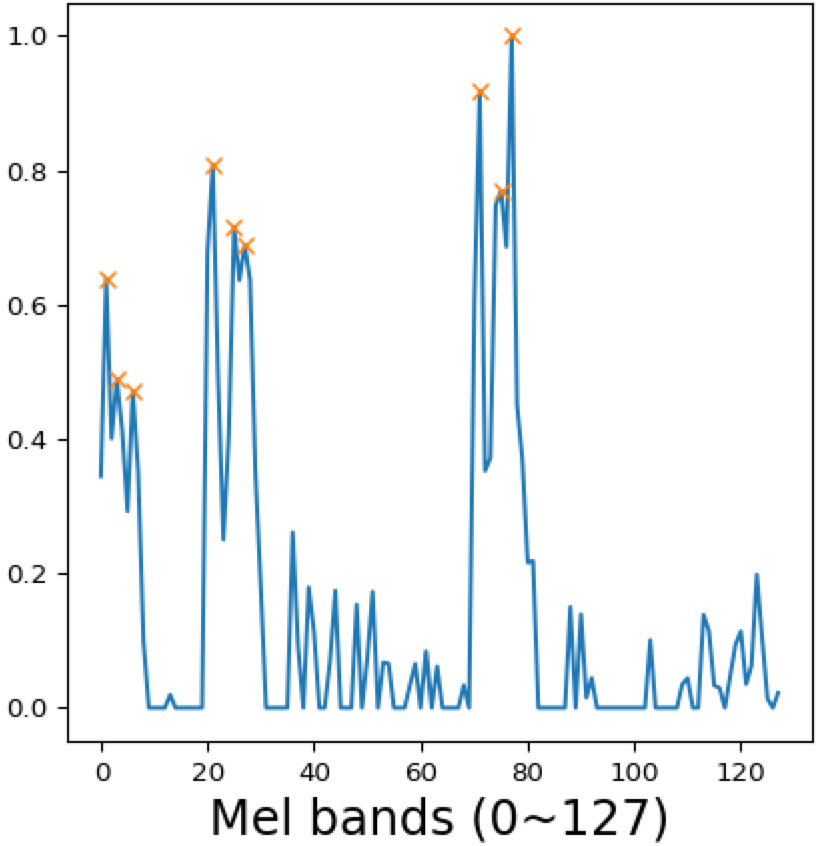




loc of Max [cossim(a_2,b_1) cossim(a_2,b_2) cossim(a_2,b_3) … cossim(a_2,b_n)]

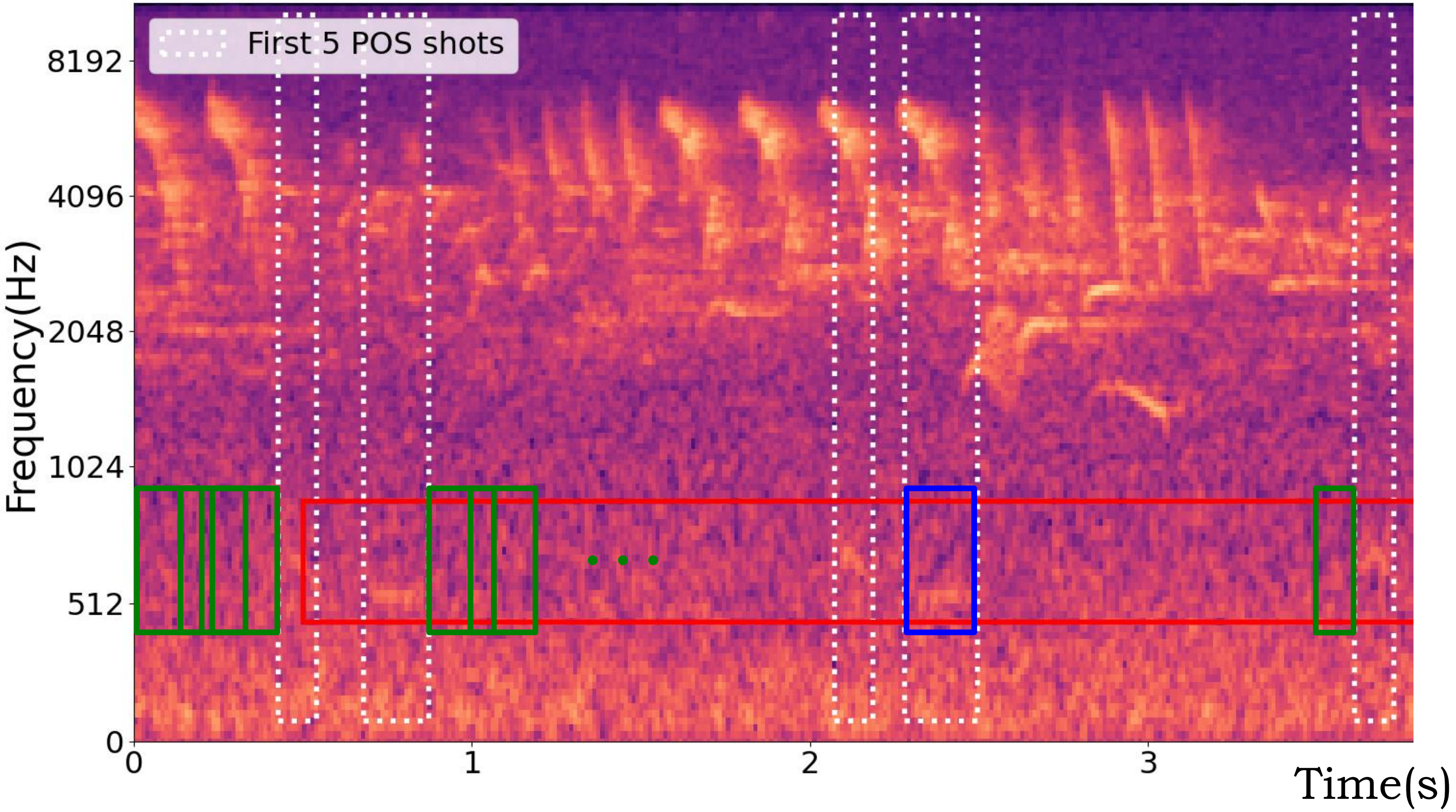
FMD(POS) – FMD(NEG)(the most similar to POS)

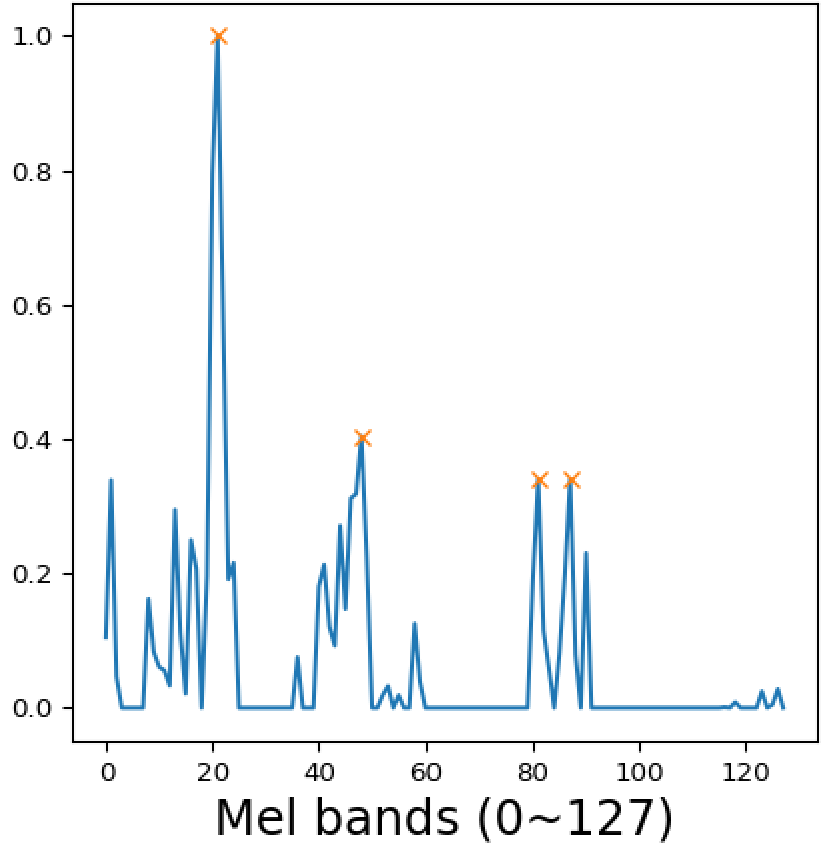




loc of Max $[\text{cossim}(a_3,b_1)\ \ \text{cossim}(a_3,b_2)\ \ \text{cossim}(a_3,b_3)\ \ \ldots\ \ \text{cossim}(a_3,b_n)]$

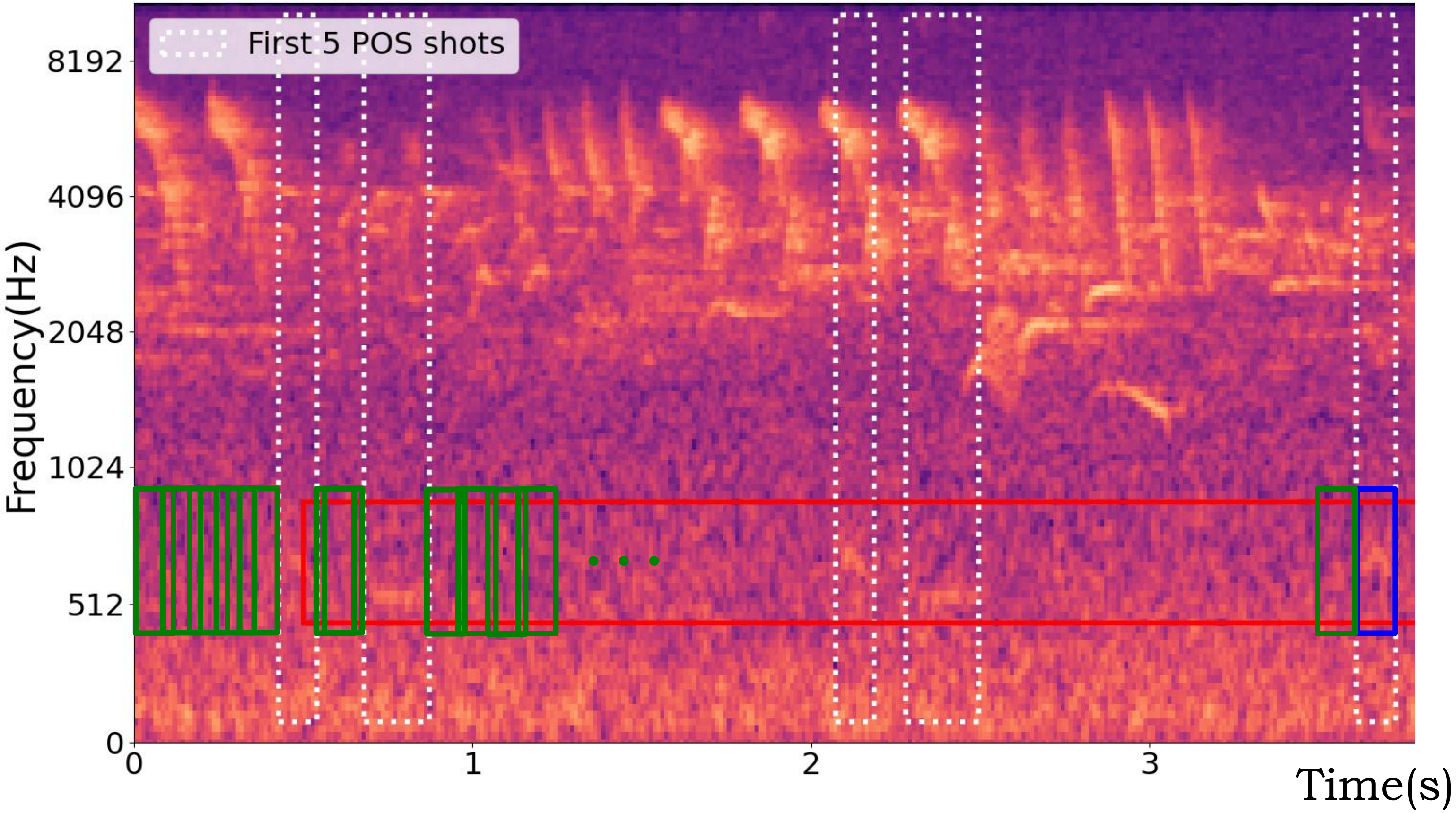
FMD(POS) – FMD(NEG)(the most similar to POS)

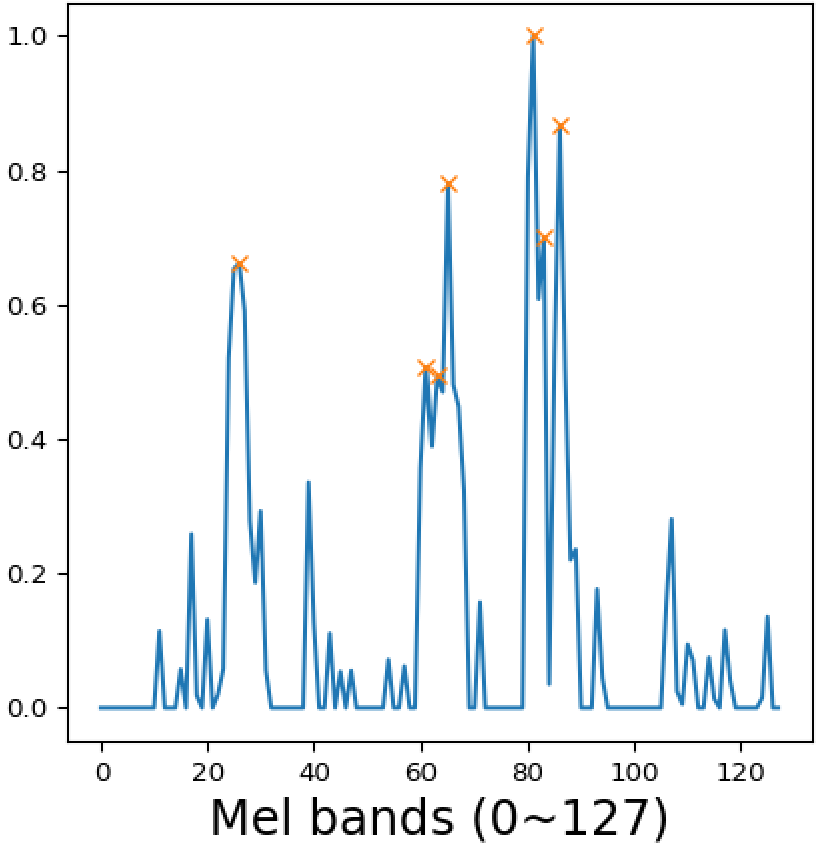




loc of Max [cossim(a_4,b_1) cossim(a_4,b_2) cossim(a_4,b_3) … cossim(a_4,b_n)]

FMD(POS) – FMD(NEG)(the most similar to POS)



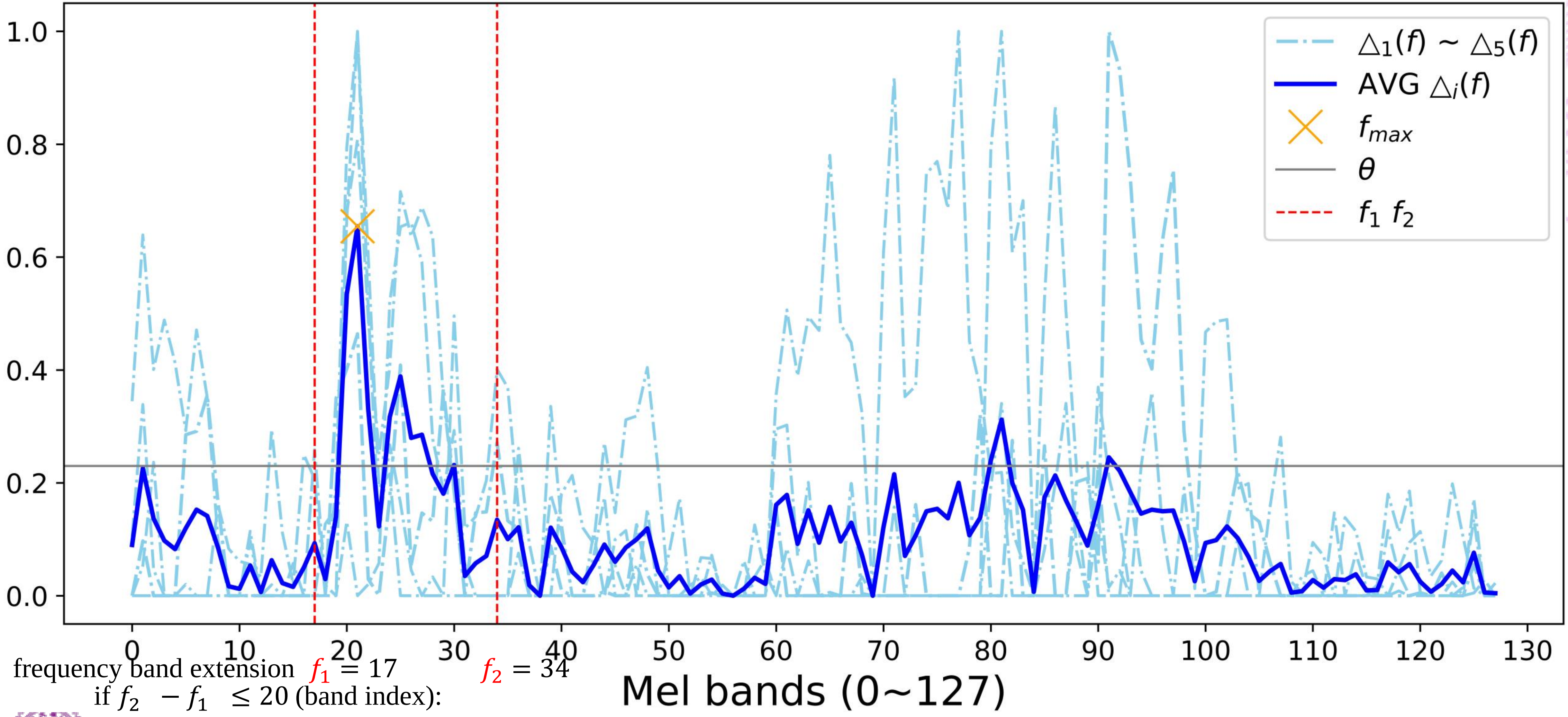


loc of Max [cossim(a_5,b_1) cossim(a_5,b_2) cossim(a_5,b_3) … cossim(a_5,b_n)]

$f_1 = 20$ ← 7ba

7ba → $f_2 = 31$

stops when all values within the zone are below θ



frequency band extension $f_1 = 17$ $f_2 = 34$

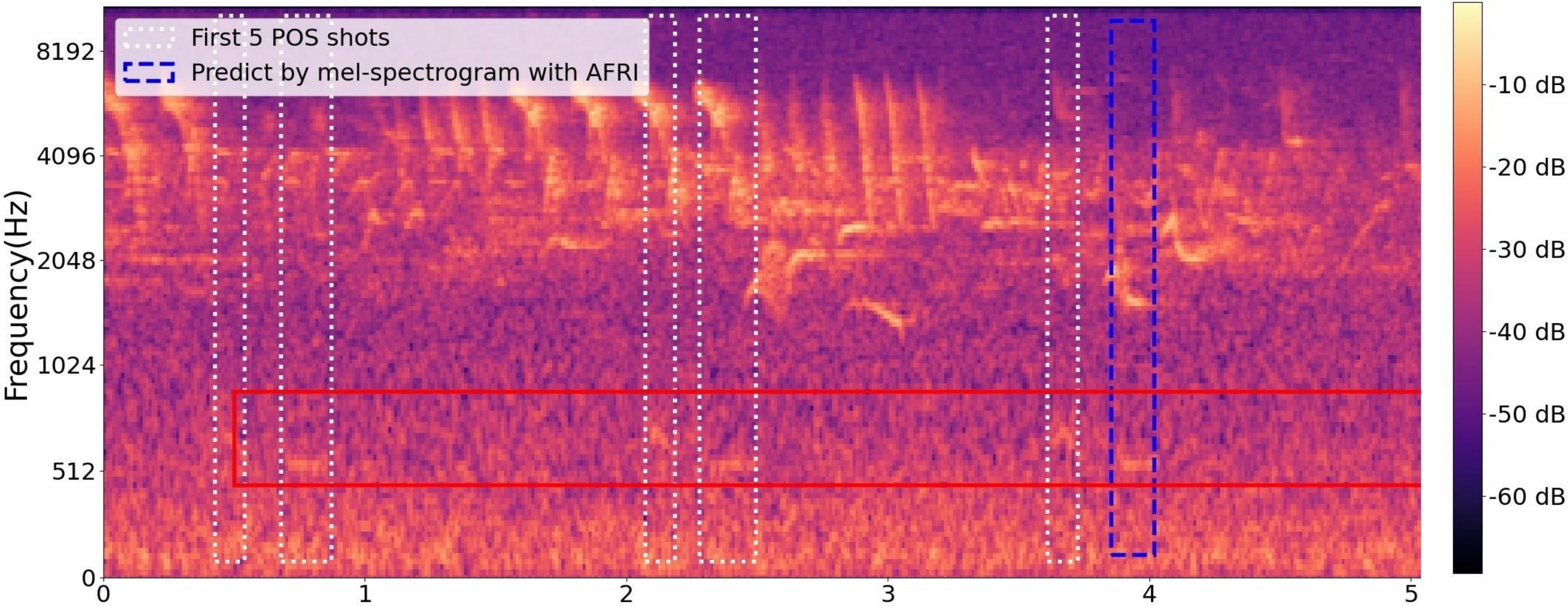
if $f_2 - f_1 \leq 20$ (band index):

f_2 is increased by 3 and f_1 is decreased by 3.

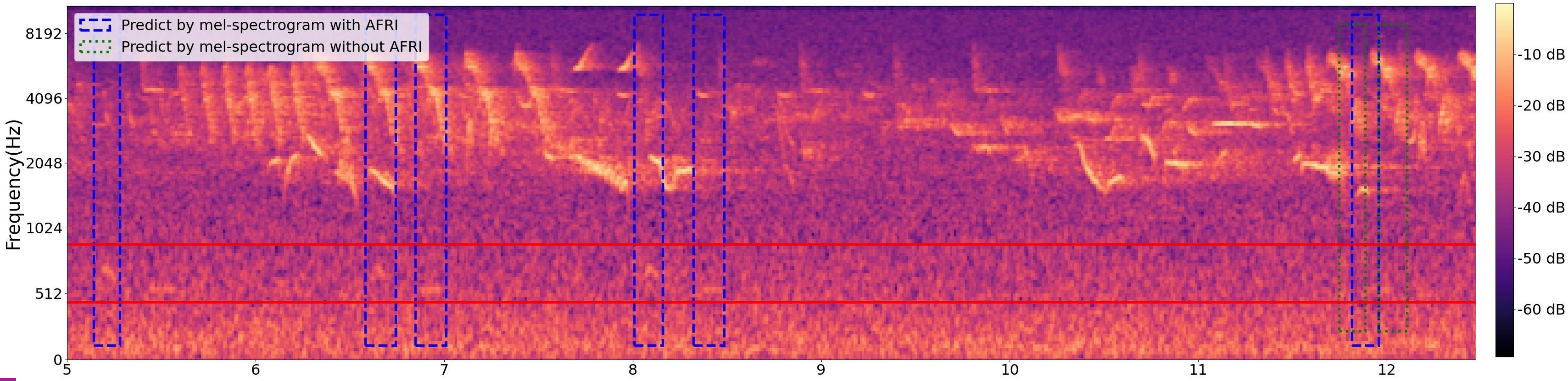
else:

f_2 is increased by 6 and f_1 is decreased by 2.

Mel-spectrogram visualization



Mel-spectrogram visualization



DCASE2024 Challenge task 5

Table 6: Results of teams' rankings on the 2024 evaluation set

System	F-score	F – CHE	F – CT	F – MGE	F – MS	F – QU	F – DC	F – CHE23	F – CW
2024 No.1 **External datasets,** **Data augmentation**	65.2 %	64.1 %	69.4 %	93.1 %	78.9 %	41.2 %	64.2 %	79.7 %	75.5 %
2024 No.2 **Data augmentation**	56.7 %	54.6 %	47.6 %	70.5 %	62.8 %	47.7 %	52.1 %	78.0 %	69.0 %
2024 No.3 **AFRI**	46.9 %	92.5 %	16.6 %	83.0 %	59.7 %	44.5 %	61.1 %	71.5 %	70.3 %
Baseline **Prototypical Network**	41.6 %	37.9 % small	28.6 % small ↑↓	41.3 % \ / overlap	58.7 % P dur	26.3 % ↑↓ overlap	47.2 % P dur overlap weak P small	68.3 % P dur overlap	62.7 % weak P