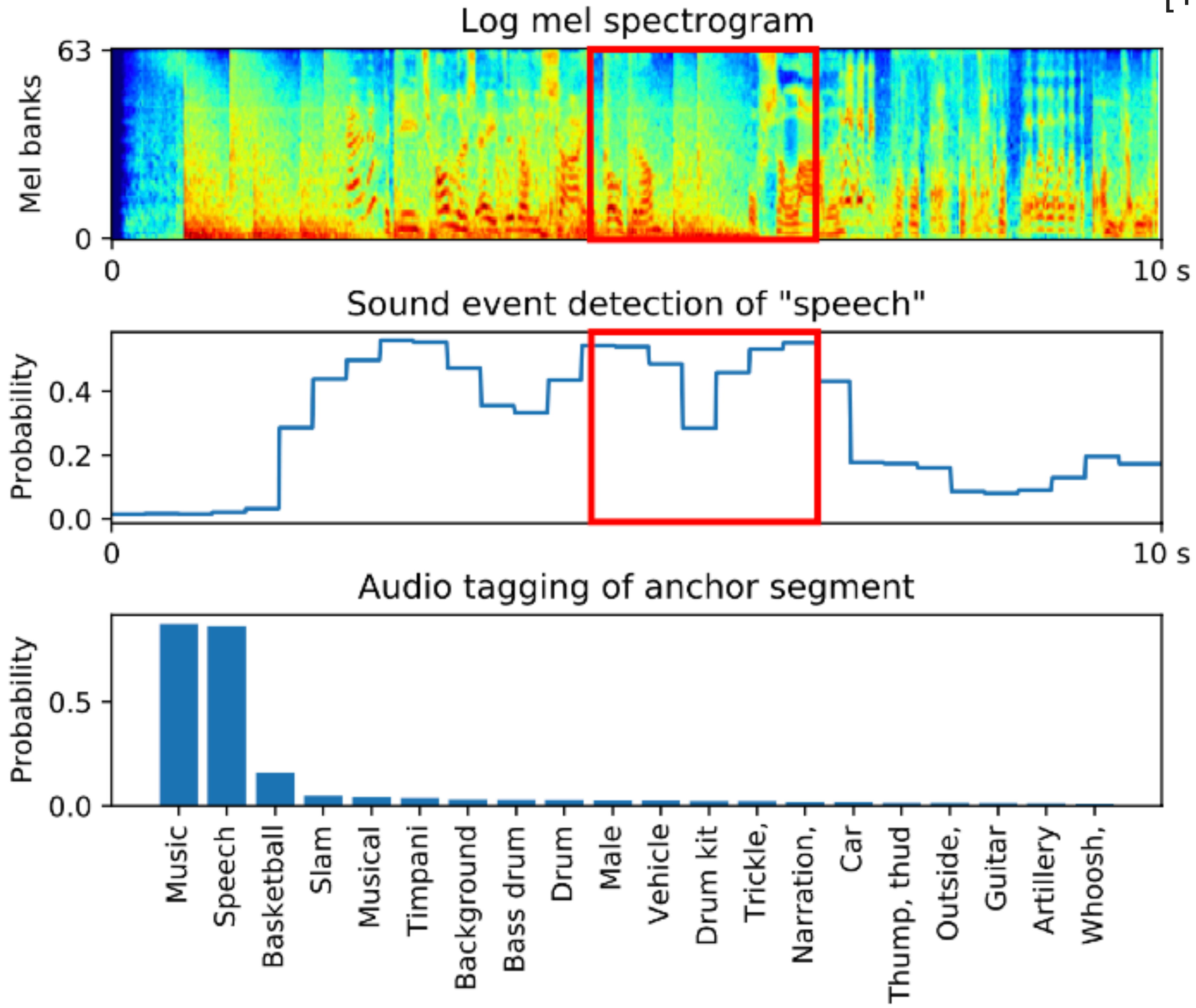
Speech Enhancement

Training phase

- Method: $f(s_1 + s_2 \mid c) \mapsto s_1$

1. The selection of s_1 and s_2.
   - s_1, s_2 should have disjoint labels.
2. The generation of condition c.
   - $c = \text{max\_pool}(f_{SED}(s_1))$
   - SED: Sound Event Detections[2]
3. The modeling of $f(\cdot)$:
   - A IRM based Conditional-UNet.

[1]



[1] Kong, Qiuqiang, Haohe Liu, Xingjian Du, Li Chen, Rui Xia, and Yuxuan Wang. "Speech enhancement with weakly labelled data from AudioSet." *arXiv preprint arXiv:2102.09971* (2021).

[2] qiuqiangkong. qiuqiangkong/panns_inference. GitHub. Published August 17, 2020. Accessed November 24, 2021. https://github.com/qiuqiangkong/panns_inference

Speech Enhancement

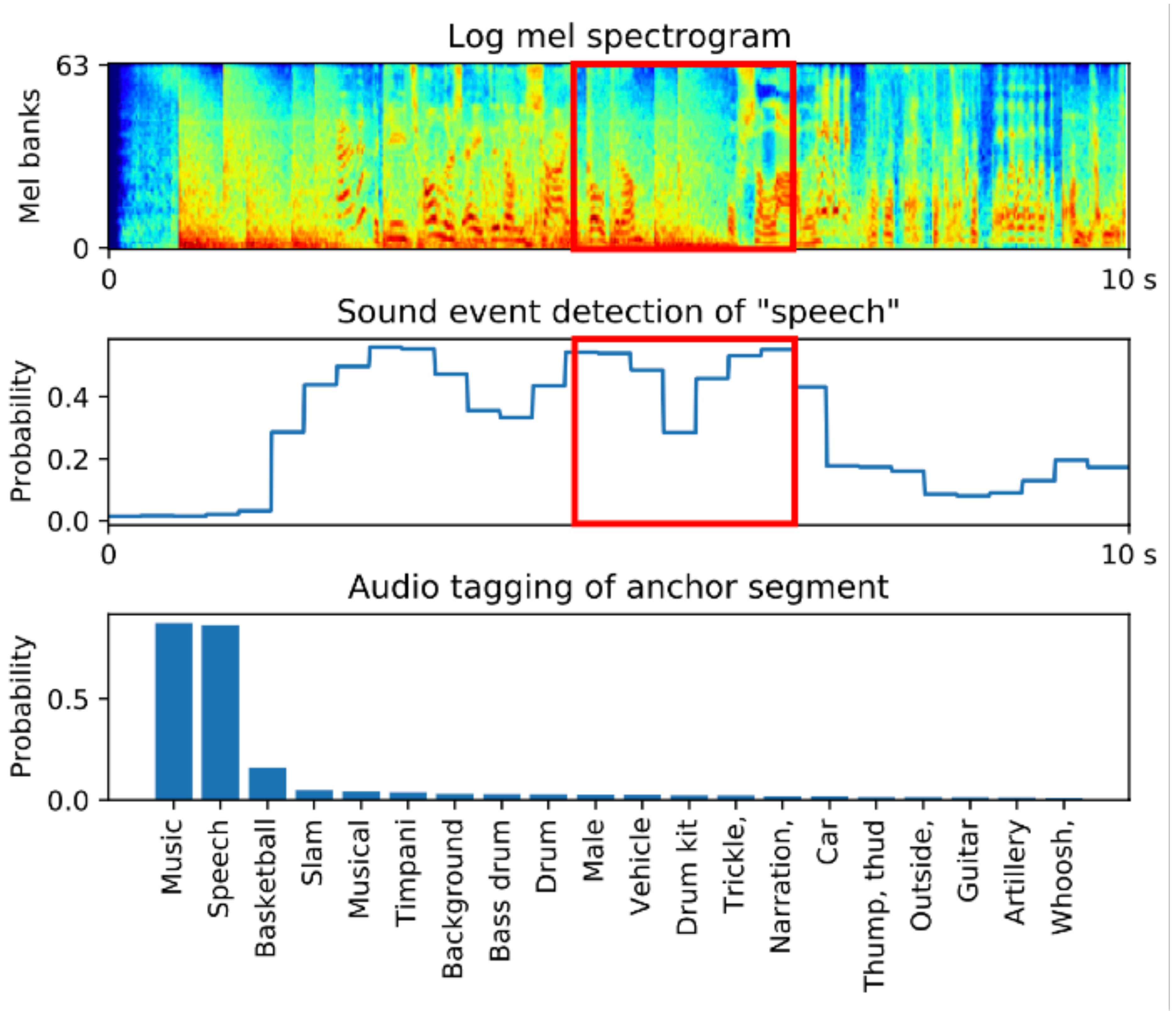
Inference phase

- Method: $f(m \mid c) \mapsto s$

1. m is the given mixture signal

2. c is the one-hot condition for speech.

3. s is the separation target.



Log mel spectrogram
Mel banks
63
0
0
10 s
Sound event detection of "speech"
Probability
0.4
0.2
0.0
0
10 s
Audio tagging of anchor segment
Probability
0.5
0.0
Music
Speech
Basketball
Slam
Musical
Timpani
Background
Bass drum
Drum
Male
Vehicle
Drum kit
Trickle,
Narration,
Car
Thump, thud
Outside,
Guitar
Artillery
Whoosh,