Utilizing Embedding Methods for Soundscape Analysis of Forest Animal Vocalization based on azimuth and elevation localization

Hao Zhao (Nagoya University), Reiji Suzuki (Nagoya University), Ryosuke Kojima (Kyoto University), Takaya Arita (Nagoya University), Kazuhiro Nakadai (Tokyo Institute of Technology)

The interest in ecoacoustics has led to an influx of novel methods for soundscape analysis. Unlike traditional automated recording units (ARUs) with a single microphone, robot audition techniques using microphone arrays can further contribute to understanding soundscape dynamics and structure in ecological environments. In addition, machine learning techniques can handle large data sets and complex soundscapes, which may offer a more efficient and accurate alternative to human analysts in soundscape analyses.

We are analyzing the early summer forest soundscape in Japan, dominated by the vocalizations of birds and cicadas, as a typical soundscape composed of multiple species. It has been reported that birds significantly avoid temporal overlap with cicadas by reducing and often shutting down vocalizations at the onset of cicada signals that utilize the same frequency range (Hart et al. 2015). We use HARKBird (Suzuki et al. 2017), a bird song localization software based on the robot audition software HARK, to conduct azimuth-elevation localization and sound separation from the recordings of a self-developed 16-channel semi-spherical microphone array, CHIRPY. However, our previously proposed classification method (Zhao et al. 2023) using acoustic indices developed for an 8-channel circular microphone array, TAMAGO, did not perform well for the elevation-azimuth localization. This is because the increased dimensions of localization space yielded false detected or low-quality reflected sounds.

This study considers extracting the vocalizations of birds and cicadas utilizing the classification methods based on recently emerging embeddings of three pre-trained models: BirdNET and wav2vec2 trained with bird songs or human speech. It was confirmed that BirdNET, pre-trained with bird songs, can effectively classify the acoustic events of other species (Ghani et al. 2023). However, it is unclear whether wav2vec2, widely used for human speech processing, can be applied to the classification of animal vocalizations. We extracted the latent vectors of separated sounds from azimuth-elevation localization results using the three models with HARKBird. We then trained a support vector machine (SVM) classifier using the latent vectors with the manually annotated labels data (i.e., bird, cicadas, noise). We applied each classifier to a ten-minute recording to extract the vocalizations of birds and cicadas, and conducted a preliminary analysis of their vocal activity and azimuth-elevation information compared with the results of manual annotation.

The classifiers using these embeddings worked better than our previous simple method. The previous classification methods based on acoustic indices revealed a low accuracy of 40%. In contrast, the classification accuracy using BirdNET reached 67%. Employing the wav2vec 2.0 model, which was pre-trained on bird songs, resulted in an accuracy of 65%, demonstrating better performance on human speech with an accuracy of 51%. This is because the false-detected sounds or low-quality sounds resembling the vocalizations of target animals are often mistakenly identified as the vocalizations of target animals by the classification based on the acoustic indices. However, the embeddings based on machine learning can preserve latent feature differences between them and target animal vocalizations, enabling discrimination between them. It also turned out that embeddings based on bird songs performed better than that based on human speech.

We further conducted preliminary analyses of the vocal activities and spatial structures of vocalizations. They demonstrated that based on the pre-trained models with bird songs, the classifiers could extract the overall soundscape dynamics and structure of birds and cicadas. Moreover, they imply there might exist temporal overlap avoidance behaviors between birds and cicadas, as expected from our previous results, constructing the distinctive spatial structure of vocalization dynamics in the azimuth-elevation space of the forest soundscape.