ABSTRACT

Analyzing music structure involves identifying and grouping similar segments to reveal an overarching structure. A key challenge in this field is capturing long-term contextual dependencies without incurring excessive computational costs. While deep learning models have shown some promising results, the computational complexity of their attention mechanisms remains a significant hurdle. Neighborhood Attention (NA) was introduced to mitigate this issue by reducing the time complexity of the attention mechanism while preserving the model's ability to recognize contextual patterns. NA has demonstrated effectiveness in music structure analysis tasks on separated audio. However, its efficiency had not been fully optimized. To address this, Fused Neighborhood Attention (FNA) was developed to reduce NA's quadratic complexity into a linear one. This research applies and investigates the effectiveness of FNA in music structure analysis of separated audio by modifying an existing NA-based model. The FNA model was compared against NA using music segmentation accuracy and computational efficiency metrics. The FNA model achieved a best F-measure Hit Rate with a tolerance of 0.5 second (HR.5F) of 0.516 and a Pairwise F-measure (PwF) of 0.611. For shorter records, FNA has shown lower inference latency compared to NA. This comparison indicates that FNA successfully enhances model performance with only a minor trade-off in accuracy.

Keywords : Fused Neighborhood Attention, musical structure analysis, music information retrieval