MIDI 2025 Conference

Deep Learning for Real-Time Sound Order Recognition in Human-Robot Interaction

Abstract

Recognizing the temporal order of overlapping sounds is an underexplored challenge in human-robot interaction (HRI), with direct relevance to applications such as first responder detection systems.

This paper presents a deep learning framework for real-time sound order recognition using recordable buzzers that emit distinct non-verbal sounds (cat meows, dog barks, helicopter noises). A multi-branch convolutional neural network (CNN) processes Mel spectrograms, Mel-frequency cepstral coefficients (MFCCs), and short-time Fourier transform (STFT) features, with an attention-based fusion mechanism to emphasize critical temporal cues. Experiments were conducted under sameamplitude, varied-amplitude, and unseen sound conditions. The proposed system achieved 99% accuracy in balanced overlaps, 91% under amplitude variation, and 74% on unseen test data with normalization. These results demonstrate that deep learning can reliably recognize sound order in overlapping conditions, supporting practical HRI scenarios. While experiments were conducted on carefully controlled synthetic overlaps, we additionally report latency benchmarks demonstrating real-time feasibility and provide an extended discussion on generalization, ecological validity, and deployment challenges in real-room environments.