【主論文の要旨・記入例】

学位報告 4

報告番号	※甲　　第　　号

主　論　文　の　要　旨

論文題目　A Study on Environmental Sound Modeling based on Deep Learning
(深層学習に基づく環境音モデリングに関する研究)

氏　　名　　林 知樹

論 文 内 容 の 要 旨

With recent increases in computational power, environmental sound understanding has become more feasible. Researchers intend to develop automated systems that can identify every possible sound in a given environment, from the sound of glass breaking to the crying of children.

One of the most important tasks in this field is sound event detection (SED), which is the task of detecting the beginning and the end of sound events and labeling them. Sound events include a wide range of phenomena that vary widely in acoustic characteristics, duration, and volume, such as the sound of glass breaking, typing on a keyboard, knocking on doors, and human speech. This diversity of targets makes SED challenging. Though recent advances in machine learning techniques have improved the performance of SED systems, various problems remain to be solved.

This thesis addresses three problems that affect the performance of monophonic, polyphonic, and anomalous SED systems. The first is how to different types of signals can be used in combination, to extend the range of detectable sound events. The second is how to model the duration of sound events, which is among the most important characteristics, to improve polyphonic SED performance. The third is how to model normal environments, in which no anomalous sound events occur, to improve the performance of anomalous SED systems.

As a part of the work in developing a life-logging system, this work focuses on the use of multi-modal signals to extend the range of detectable sound events into the area of common human activities. The key to realizing these applications is finding ways to associate different types of signals to detect variety of human activities. First two deep neural network (DN

N)-based fusion methods using multi-modal signals are proposed for this associative goal. Then, a large database of human activities recorded under realistic conditions is created for testing the performance of the proposed methods. Furthermore, to address the problem of model individuality, which degrades system performance, speaker adaptation techniques from the field of automatic speech recognition are introduced. Experimental results using the constructed database demonstrate that the use of multi-modal signals is effective, and that speaker adaptation techniques can improve performance, especially when using only a limited amount of training data.

To improve the performance of polyphonic SED systems, this work focuses on modeling the duration of sound events. To do this, a novel hybrid approach using duration-controlled long short-term memory (LSTM) is proposed, which builds upon a state-of-the-art SED method that performs frame-by-frame detection using a bidirectional LSTM recurrent neural network (BLSTM) by incorporating a duration-controlled modeling technique based on a hidden Markov model (HMM) or a hidden semi-Markov model (HSMM). The proposed approach makes it possible to model the duration of each sound event precisely and to perform sequence-by-sequence detection without needing thresholding.

Furthermore, to effectively reduce insertion errors, post-processing method using binary masks is also introduced. This post-processing step uses a sound activity detection (SAD) network to identify segments for activity indicating any sound event. Experimental evaluation with the DCASE2016 task 2 dataset demonstrates that the proposed method outperforms conventional polyphonic SED methods, proving that sound event duration is effectively modeled for polyphonic SED.

The key to successful anomalous SED is in finding a method for modeling the normal acoustic environment. This modeling is performed conventionally in the acoustic feature domain, but this results in a lack of information about temporal structure like the phase of the sounds. To address this issue, a new anomalous detection method based on WaveNet, which is an autoregressive convolutional neural network that directly models acoustic signals in the time domain, is proposed. The proposed method uses WaveNet as a predictor rather than as a generator to detect waveform segments that are responsible for large prediction errors as unknown acoustic patterns. Furthermore, to consider differences in environmental situations, i-vector is utilized as an additional auxiliary feature of WaveNet. The i-vector extractor should allow the system to discriminate the sound patterns, depending on the time, location, and surrounding environment. Experimental evaluation with a database of sounds recorded in public spaces shows that the proposed method outperforms conventional feature-based approaches, and that time-domain modeling in conjunction with the i-vector extractor is effective for anomalous SED.