

論文 / 著書情報
Article / Book Information

題目(和文)	深層学習ベースの環境音セグメンテーション-音源定位、分離、クラス分類の統合-
Title(English)	Deep-Learning-based Environmental Sound Segmentation – Integration of Sound Source Localization, Separation, and Classification –
著者(和文)	周藤唯
Author(English)	Yui Sudo
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第11924号, 授与年月日:2021年3月26日, 学位の種別:課程博士, 審査員:田中 正行,蜂屋 弘之,大山 真司,原 精一郎,宮崎 祐介
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第11924号, Conferred date:2021/3/26, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	要約
Type(English)	Outline

Doctoral Dissertation

**Deep-Learning-based
Environmental Sound Segmentation
–Integration of Sound Source
Localization, Separation,
and Classification–**

Yui Sudo

Department of Systems and Control Engineering,
School of Engineering,
Tokyo Institute of Technology

December 2020

Abstract

This research aims to understand various environmental sounds, such as sound source signals and class of sound sources, from signals observed by the microphones, with the aim of high level human-robot interaction. Specifically, this research focuses on “environmental sound segmentation” for active section detection, sound source localization, sound source separation, and classification for various kinds of environmental sounds.

In recent years, there has been a lot of research on automatic speech recognition and speech synthesis to enable human-robot interaction. For real-world applications, robust methods are required due to noise and multiple acoustic events. Various methods such as sound source localization, sound source separation, and classification have been proposed in acoustic signal processing, robot audition, and machine learning for use in real-world environments containing multiple overlapped sound events.

Traditional approaches use the cascade method, incorporating individual functions based on array signal processing techniques. The main problem in these methods is accumulation of errors generated at each function. Because each function is optimized independently regardless of the overall task, each output might not be optimal for subsequent blocks.

Recently, deep-learning-based end-to-end methods using a single-channel microphone that does not use a conventional cascade system have been proposed. These studies have made it possible to perform multiple functions, such as section detection, sound source localization, sound source separation, and classification, simultaneously on a single network. Most previous studies, however, have dealt with only two classes of sounds, speech and others, with few studies targeting multiple classes of sounds such as environmental sounds. In performing more general-purpose interactions, robots must not only separate human voices

but environmental sounds, which include many classes. These methods dealing with human voices can be extended to environmental sounds containing multiple classes. The duration of sound events and those frequency components, however, may vary greatly from class to class, which may degrade performance. Additionally, performance deteriorates with overlapped sounds from multiple sources, because a single-channel microphone does not exploit spatial features.

Multichannel-based methods have been proposed. These methods can improve the performance of sound source localization and sound source separation by utilizing spatial features depending on the direction of arrival. However, in addition to sound source localization and sound source separation, it is necessary to perform classification by utilizing spectral features in the environmental sound segmentation. Normally, the direction of arrival and the class are not correlated each other because the relative angles of the sound sources and microphones are not fixed. Therefore, in environmental sound segmentation that uses these features simultaneously, the network overfits to the relationship between the direction of arrival and the class.

This paper proposes an deep-learning-based environmental sound segmentation method that integrates section detection, sound source localization, sound source separation, classification to prevent error accumulation as in traditional cascade methods. In addition, we propose a method for dealing with spectral and spatial features simultaneously in multichannel environmental sound segmentation.

Chapter 1 describes the motivation, objectives of this study, and technical issues of related works, and then describes the structure of this dissertation.

In Chapter 2, we first define the environmental sound segmentation, which deals with section detection, sound source localization, sound source separation, and classification in an integrated manner. Related works on environmental sound segmentation are divided into two approaches: the cascade approach and the integrated approach, and each of these issues is described. After that, we explain the positioning of this dissertation.

Chapter 3 describes a single-channel environmental sound segmentation that integrates section detection, sound source separation and classification. We proposed a method that combines sound event detection and semantic segmentation methods, which is robust to different sound event lengths across classes. The effectiveness of the integrated method in preventing the accumulation of errors was evaluated by comparing it with a traditional cascade approach using array signal processing techniques. The robustness of the proposed method to environmental sounds with various event lengths and spectrum was then tested by comparing the proposed method with conventional single-channel methods. Simulation experiments using the created datasets including 75-class environmental sounds showed that the proposed method can prevent the error accumulation of traditional cascade methods and is robust to different event lengths.

Chapter 4 describes the method using multichannel microphones. Using multichannel microphones, spatial features such as inter-channel phase difference between microphones can be used to estimate the direction of arrival and to separate the sound sources coming from different directions. However, periodic functions, such as inter-channel phase difference and the direction of arrival, have discontinuity points and therefore learning can be unstable when used in neural networks. Chapter 4 focuses on the task of sound event localization and detection and proposed a technique to handle periodic functions such as the inter-channel phase difference and direction of arrival of sound sources efficiently in a neural network. The performance of the proposed method was evaluated in comparison to conventional methods using the periodic function and it was shown to reduce the estimation error of the direction of arrival.

In Chapter 5, we proposed a multichannel environmental sound segmentation method that does not overfit to the relationship between the direction of arrival and the class. This method prevents overfitting to the relationship between the direction of arrival and the class by explicitly separating the sound source localization and separation block from the sound source separation and classification

block. Simulation experiments using the created datasets including 75-class environmental sounds showed that the proposed method improved the segmentation performance compared to conventional methods. When the sound sources arrived from a same direction, the sound source separation and classification block was able to correct the sound sources that could not be separated by the sound source localization and separation block. In other words, the proposed method was able to efficiently exploit spectral and spatial features while preventing the accumulation of errors in the sound source localization and separation block.

Chapter 6 summarizes and concludes this dissertation.