Leveraging User Input and Feedback for Interactive Sound Event Detection and Annotation

Bongjun Kim
Northwestern University
Evanston, IL, USA
bongjun@u.northwestern.edu

ABSTRACT

Tagging of environment audio events is essential in many areas. However, finding sound events and labeling them within a long audio file is tedious and time-consuming. Building an automatic recognition system using modern machine learning is often not feasible because it requires a large number of human-labeled training examples and it is not reliable enough for all uses. I propose interactive sound event detection to solve the issue by combining machine search with human tagging, specifically focusing on the effectiveness of various types of user-inputs to the interactive sound searching. The types of user inputs that I will explore include binary relevance feedback, segmentation, and vocal imitation. I expect that leveraging one or combination of these user inputs would help users find audio contents of interest quickly and accurately, even in the situation where there are not enough training examples for a typical automated system.

Author Keywords

interactive machine learning; sound event detection; human-in-the-loop system

INTRODUCTION

Searching for sound events within a long audio file can be used in many areas. Sound designers might want to search for sound effects in a long audio recording quickly. Ecologists need tools for labeling bird calls and singing in lengthy recordings (e.g. a 24 hour long recording of a natural scene). Police officers might want to find suspicious sound events in city recording. Speech and language pathologists often wish to label sound and speech events in day-long (24 hours) recordings of patient environments. Collecting audio segments of interest is also an inevitable step when ones want to create an audio dataset from their own recordings, which can be used as sound design library or training dataset for machine learning. The current standard way of searching for audio is text-based search. However, text-based search within a lengthy recording is currently possible only when someone has manually added time-coded text, which is prohibitively labor intensive.

An obvious solution to the issue is to use an automated audio retrieval system. However, building a fully automated system leveraging machine learning techniques such as deep learning [4] typically requires a large number of training examples (e.g. hundreds of thousands of examples) during system design. On top of that, the set of classes to be identified must be known prior to deployment, meaning end-users cannot define new classes to be identified. Even with lots of training data and model tuning, machine labeling may not be good enough for some applications that require human-level accuracy.

I propose a mixed-initiative audio search system that is useful in a situation where there are no time-coded text labels available in the audio and there are too few training examples to successfully train a statistical machine learner, yet the search task is too time consuming to perform entirely manually. The proposed system allows a user to define a target sound class on-the-fly and helps a user to find the target sound events within a long audio file more quickly than is possible by manual search. The goal of the retrieval system is to speed up the labeling task at hand, rather than to train a generalized machine learning model for later use on different data.

RESEARCH GOALS

I apply a human-in-the-loop approach to content-based audio search. The idea is to engage a user in an interactive process to collaboratively search for sound with machines. The success of interactive retrieval depends on how to incorporate human knowledge into the system. A user can guide and improve the system by providing it with information about a target concept to search for. My research goals are following: 1) Developing new user input and feedback methods to provide machines with more information about a target sound, 2) Implementing a sound event annotation interface that embodies all the interaction methods.

BACKGROUND AND RELATED WORK

Interactive machine learning has been applied to document and image retrieval systems [1, 5], where a user can review retrieved documents or images and provide feedback. This approach is aimed at training the best classifier to retrieve items relevant to a query. My goal is to speed up the labeling task at hand, rather than to train a generalized machine learning model for later use on different data.

IUI 2018, March 7–11, 2018, Tokyo, Japan.
ACM ISBN 978-1-4503-4945-1/18/03.
https://doi.org/10.1145/3172944.3173149

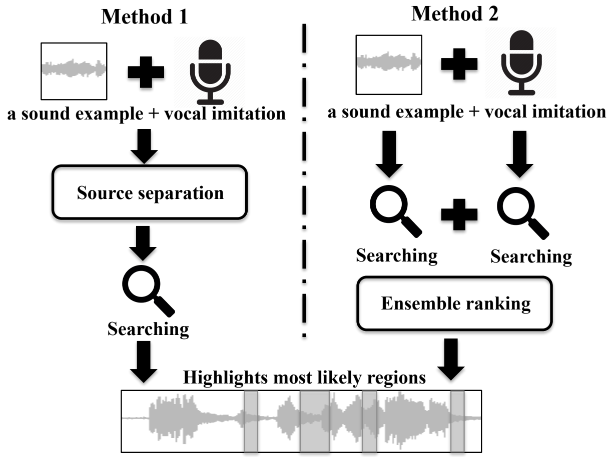
Method 1
a sound example + vocal imitation
Source separation
Searching
Method 2
a sound example + vocal imitation
Searching
Searching
Ensemble ranking
Highlights most likely regions

Figure 1. vocal imitation processing methods

How to leverage the user input and feedback in an interactive learning interface could vary depending on the types of data to search (e.g. text, image, or sound). There are only a few works that have applied interactive learning to audio retrieval system [2] and they did not address the difference between audio and other types of data in terms of possible user-inputs for interactive audio retrieval. My focus in this work is how to leverage user input (i.e. human knowledge) in interactive audio retrieval and how much the interaction would help a user to achieve their goal quickly and easily.

CURRENT RESEARCH STATUS

The proposed interface will allow three different types of user inputs: binary relevance feedback (positive or negative), user segmentation (adjusting time-boundaries of retrieved sound segments), and vocal imitation (imitating sound with vocalization). I have implemented my initial proof-of-concept for the proposed system [3] that allows only relevance feedback and user segmentation. How the system works is following. Given a long audio file, a user selects a region that contains a target sound event in the audio track as an initial search key. The system returns the *n* most similar sound events. Then, the user gives feedback to the system and the importance of audio features is re-weighted so that it can return more relevant examples in the next round. Readers can watch the demo video at http://www.bongjunkim.com/ised/. The user study showed that the mixed-intuitive approach speeds up a sound annotation task. I have also learned that, while the interface improves the overall performance of human's audio searching, the human-machine interaction loop causes extra interaction overhead. I will redesign the human-machine interaction to reduce the extra cost during the next iteration of the research.

METHODS AND RESEARCH PLANS

My prior works use two types of user feedback: 1) applying positive or negative labels to each suggested region, 2) adjusting boundaries of the suggested region when its onset and offset are not perfectly aligned with the target sound.

On top of the interaction methods, I plan to develop vocal imitation-based interaction. Vocal imitation is a simple and effective way to describe sound concepts. Unlike texts and images, audio events are often fully overlapped (e.g. a cough is concurrent with television noise). Unless a user tells a system which sound event is the target, the retrieval results would be degraded. In this work, a user will be able to specify the target sound by imitating it. I am exploring two possible approaches to processing the vocal imitation as shown in Figure 1: 1) extracting target source by vocal imitation, 2) combining the two search results from an actual example and its vocal imitation.

I also plan to implement an interface for sound event detection that embodies all the proposed user input and feedback method to speed up the sound labeling task. Based on participants' feedback from the initial experiment [3], I will redesign my initial interface and add a new component for vocal imitation-based interaction where users can quickly monitor or edit their vocal imitation query and feedback.

CONCLUSION AND INTELLECTUAL MERIT

Developing interaction methods that help to speed up sound event search are the main advances of this work. Vocal imitations of sound events would help machines understand perceptually relevant features of audio, and leveraging a combination of various types of user inputs would help users find audio contents of interest quickly and accurately even in the situation where there are not enough training examples for an automated system. I expect that the proposed work will provide insight about how human input and feedback can reduce the gap between human and machine's auditory perception and how they can be applied to an intelligent interactive system.

ACKNOWLEDGMENTS

This work was supported by NSF Grant 1617497.

REFERENCES

1. Saleema Amershi, James Fogarty, Ashish Kapoor, and Desney Tan. 2011. Effective End-user Interaction with Machine Learning. In *Proc. of the AAAI Conference on Artificial Intelligence (AAAI)*. AAAI Press, 1529–1532.
2. Sébastien Gulluni, Slim Essid, Olivier Buisson, and Gaël Richard. 2011. An interactive system for electro-Acoustic music analysis. In *Proc. of the International Society for Music Information Retrieval Conference (ISMIR)*. 145–150.
3. Bongjun Kim and Bryan Pardo. 2017. I-SED: An Interactive Sound Event Detector. In *Proc. of the International Conference on Intelligent User Interfaces (IUI)*. ACM, 553–557.
4. Giambattista Parascandolo, Heikki Huttunen, and Tuomas Virtanen. 2016. Recurrent neural networks for polyphonic sound event detection in real life recordings. In *Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6440–6444.
5. Burr Settles. 2011. Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. In *Proc. of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 1467–1478.