Context-based sound event recognition
Niessen, Maria Elisabeth

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2010

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
The research domain of automatic sound event recognition aims to describe an audio signal in terms of the sound events that compose a sonic environment. The ability to recognize events in a real-world environment requires a listener or system to separate the sound events from each other and the background. Furthermore, these separated events need to be recognized. To recognize a sound event implies that some representation of the event is already known to the receiver, and can be identified when it is encountered again. The ability of sound event recognition depends not only on the audio pattern specific to the event, but also on the semantics of the event. For example, a sound of a purring cat may seem unique, but without any other information than represented by the audio signal, it can sound like an engine as well. In this thesis we show that the task of recognizing a sound event can be alleviated with the semantics of the event, which is inferred from a model of the context in which the event occurs.

A possible strategy to provide an automatic sound recognition system with the semantics of a sound event, is to develop it for a specific application, and hence a specific type of context. For example, automatic speech recognition systems expect a speech signal as input, which is ensured by a user. Therefore, particular assumptions about the audio signal can be made, and context information in the form of grammar rules can be applied to recognize a word sequence (section 2.2). However, if a system for automatic sound event recognition is not designed for a specific application, and should work in variable and uncontrolled real-world environments, no assumptions about the environmental conditions can be made. Therefore, additional analysis of the audio signal is required. First, the sound events to be recognized have to be separated from the background, because the input signal consists of more than one type of sound (section 2.3). Second, the operating environment cannot be controlled, hence the system has to deal with transmission effects, such as reverberation (section 2.4).
In addition to handling the challenges of real-world environments with signal-driven methods, the semantics of the event and its environment are essential to recognize sound events in an unreliable or ambiguous audio signal. People have no difficulty in recognizing sound events in many different and noisy situations. Hence, we developed a model for automatic sound event recognition that is inspired by the strategies of human listeners as investigated by (psycho-)acoustics and cognitive psychology. The human percept of a sound event, referred to as an auditory event (section 3.2), has properties that a representation of a sound event in an automatic system can benefit from as well. People can generalize auditory events over different experiences, environments, and senses. Therefore, our model should store invariant representations of sound events, so that it is robust to variable environmental conditions, similar to human perception. Moreover, people benefit from information about the environmental context to recognize sound events (section 3.3). This facilitatory effect of contextual knowledge is an important design objective of our model.

Some other studies have been aimed at modeling context awareness in acoustics (section 4.1). However, in these studies the goal has been to estimate the context in itself, rather than to use context for the improvement of sound event recognition. In other research domains, such as information retrieval and handwriting recognition, context has been used to improve recognition or retrieval of objects. Often methods in these research domains use spreading activation networks, which are based on a model of human memory (Collins and Loftus, 1975). In this thesis we show that spreading activation networks can also be applied to estimate the most likely interpretation of an audio pattern. We introduce a context model in which the semantics of the events and the context are represented as nodes in the network (section 4.2). The activation of the nodes spreads through the network to determine the confidence of possible interpretations of an audio pattern (section 4.3).

The advantage of modeling context in automatic sound event recognition is demonstrated by applying an integrated system to audio recorded in real-world environments. This integrated system is a combination of a signal-driven analysis of the audio signal, which provides hypotheses of sound events, and the context model, which interprets these hypotheses. Knowledge about the environmental context is learned in a training phase. This knowledge is represented as a network
of nodes, in which the nodes represent the semantics of the sound events and the different contexts. Furthermore, the connections between the nodes carry weights that indicate the probabilities that these sound events and contexts are encountered subsequently or concurrently. The values of the weights between nodes are learned from annotated training data (section 5.2). The type of context that is learned depends on the application domain and the data set. In a stable environment, information of co-occurring events can help to form expectancies of future events. For example, at a train station the beeping of a closing door is likely to be followed by a departing train. Alternatively, if the data set is recorded at qualitatively different types of locations, the estimated location can help to predict the types of sound events that may be heard. For example, birds are more commonly heard in parks than near busy roads. We explored the benefit of both types of context on sound event recognition in two experiments (section 5.3 and 5.4). The results of these experiments show that the evaluation of contextual knowledge improves the recognition of sound events compared to an exclusively signal-driven method.

Contextual knowledge is not restricted to knowledge that can be derived from the audio signal and annotations of the audio signal. Other types of knowledge can be beneficial for sound event recognition as well. Knowledge inferred from different types of input, such as sound, image, and location, can reinforce each other to obtain an increased awareness of events in the environment. Ideally, these different informational resources can be combined in a single system. Because the nodes in the context model are not described by modality specific knowledge, they can be used for other types of information. We tested the applicability of the context model on the recognition of ambiguous visual information, in the domain of robot localization. Visual information received by a robot is often ambiguous (similar to acoustic information), because similar observations, such as (parts of) chairs or windows, can be made at distinct places in an environment. Learned knowledge about the environment, and the robot’s hypothesized position in the environment, can help to disambiguate these observations. As a result, the position prediction improves compared to a signal-driven approach (chapter 6).