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).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license. More information can be found on the University of Groningen website: https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment.

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.
1

General Introduction
A simple game, named “guess the sound”, has been successful on the radio for several decades. Radio listeners have to guess the source of a short audio fragment, such as a match box being opened, shuffling playing cards, opening an umbrella, and so forth. Usually a prize can be won by giving the correct answer. For example, on a Dutch radio channel the value of the prize that can be won increases as the number of failures at guessing the audio fragment increases. Furthermore, the radio channel gives hints that should make guessing the sound easier, such as “you hold it in your hands” in case of shuffling playing cards. The difficulty of the game can be demonstrated by an analogy in the visual domain: try to recognize what the object that is depicted in Figure 1.1 represents. Why can money be won with a task that people perform effortlessly in their everyday life? The answer is related to the hints that the radio channel gives out: people use the context in which a sound is heard to identify it. When they hear sounds (or see objects) in their everyday life, they use their knowledge of the environmental context to infer which events they are likely to hear, and to discard interpretations that are unlikely given the context. Moreover, people do not only hear, they also see, smell, and feel. Therefore, they are normally not as clueless as the radio listener who tries to guess the sound.

Current automatic sound recognition systems are faced with the same problem as the radio listener. The system is presented with a short audio fragment, and has to recognize the source of the sound. The task is usually even more complicated, because the input cannot be controlled if the audio signal is recorded in a real-world environment. As a consequence, the audio signal can comprise sound produced by multiple sources. In this thesis we will demonstrate that the task of the automatic sound recognition system can be alleviated in the same manner as for the radio listener, namely by giving it knowledge about the context of the sound. To accomplish this goal we will first explore the problem of recognizing sound events in a real-world environment by reviewing strategies of both automatic systems and human listeners. Subsequently, we introduce and test a model that incorporates context into an automatic sound recognition system.

1 Whenever we introduce a term that is important in this thesis, it is printed in italic. The definitions of these terms can be found in the glossary (appendix A).
**Figure 1.1:** A visual analogy of “guess the sound”. The image is a two-dimensional schematic representation of a visual landscape, like the audio sample is stereo playback of a sonic environment. Because the context information is removed, it is difficult to guess what the image represents. Interpretation is easier when hints are given about the context of the image. For example, this image is viewed from the top, and water is surrounding the depicted object.

**1.1 Automatic Sound Recognition**

The research domain of automatic sound recognition aims at describing an audio signal in terms of the sound events or sources that compose a sonic environment. It has important (future) applications in fields as diverse as monitoring sonic environments, robotics, security systems, content-based indexing of multimedia files, and human-machine interfaces. Most sound recognition research is aimed at improving one of these application domains, such as speech recognition (O’Shaughnessy, 2008) or music genre classification (Aucouturier and Pachet, 2003). The methods in these application domains have proven successful in problems with a single, known type of sound, and even in recognizing isolated environmental sounds, such as footsteps or jangling keys (Cowling and Sitte, 2003). However, these methods have some attributes that make them less suitable for automatic sound event recognition in real-world environments.
To recognize a sound event implies that it is already known to the receiver. Therefore, a representation of the sound event needs to be stored, so it can be retrieved when an instance of the sound event is encountered. In automatic sound recognition systems this representation is typically stored as a set of temporally ordered features that describe the whole spectrum of each time frame (short time interval) in the signal—a spectrum is a representation of the energy contribution for all frequencies in one frame of an audio signal (Figure 1.2). Intervals of an audio signal or pre-selected samples are classified based on these features. However, in an audio signal of a real-world environment these intervals or samples do not necessarily correspond to (part of) a single event, because the amount of sound events contributing to the audio signal is not controlled. Multiple sound events can co-occur, at other moments no sound events occur at all. Furthermore, the audio signal can be masked or distorted by transmission effects such as background sounds and reverberation. As a consequence, a system that has to recognize sound events in a real-world environment cannot rely on the assumption that the input consists of a single, known, and undistorted signal type, as the methods used on speech and music often can.

A system for real-world sound recognition needs to segregate possible sound events from a background before it can recognize the events. Segregation of a sound event means that its constituent components in the time-frequency plane—a representation that shows the temporal development of frequency components in an audio signal (see appendix B)—are selected and grouped. As a result, correctly segregated events can be analyzed as individual elements of a sonic environment. Figure 1.3 shows a schematic example of selection and grouping in a time-frequency plane. Even if sound events are segregated and undistorted, different events can share similar audio patterns (combination of components) while they convey a distinct meaning, such as screaming and laughing. In other words, the event that produces the sound can be ambiguous, similar to the fragments in “guess the sound”. In conclusion, we need two constituents to make an advance in real-world sound event recognition. First, the segregation of audio patterns from a signal should provide hypotheses about the sound events producing these patterns. Second, we need a model that interprets these hypotheses based on contextual knowledge, and disambiguates sound events if necessary. The second task is the focus of this thesis.
Figure 1.2: One time frame of 25 milliseconds (a) and the energy spectrum of this frame (b) from an audio signal of the sound of a siren with rain and thunder in the background. The siren produces a tonal sound with a frequency around 1.1 kHz, which can be calculated from the fast zero crossing in the time domain signal, and it can be seen as a peak in the energy spectrum. Furthermore, a slower wave can be seen in the time domain signal, which is also visible as a peak around 140 Hz in the spectrum. This frequency component is caused by the thunder. The rain is a noise-like sound. Hence, its spectrum covers a broad frequency range. A feature that describes the whole spectrum includes information about all sound events that are present in the audio signal (the siren, thunder and rain).
Figure 1.3: The time-frequency plane of a recording at a square in a town (with sound events like speech, birds, and a plane), computed with a gammachirp filter bank (Irino and Patterson, 1997). The gray-scale indicates the energy in decibels (dB): darker gray corresponds to more energy. The frequency axis is logarithmic. The dashed boxes indicate possible ways to select and group events in the plane based on their local properties. For example, between approximately 1 and 3 seconds speech can be seen, which can be grouped because its components are harmonic.

1.2 HUMAN SOUND PERCEPTION

While present-day automatic sound recognition is often designed for specific tasks or specific environments, human sound perception functions for all sounds and is robust to many different environmental conditions that influence the audio signal. People can recognize a driving car, whether they hear it on a road or on gravel, in the rain or in a tunnel. Even when people have not heard a sound event before, they are able to hypothesize as to the source of the event. This ability does not rely only on auditory processing, but includes many cognitive functions as well. Depending on factors like their goals, expectations, memory, preferences, and current situation, people perceive the sound events in their environment differently.
For example, in a familiar environment people do not have to identify common sound events, because they are expected, and do not provide any new information (Grossberg, 1980).

One important factor that allows people to hear in unconstrained environments is their knowledge of the context, which helps them to form predictions and guide their perception of the environment (Bar, 2007). Events or objects in the real-world usually do not occur in isolation, but are related to other events and can be heard in particular environments (Oliva and Torralba, 2007). Therefore, the meaning of a sound event pertains to the associations that people have to other events and environments. Especially when an audio signal is unreliable or can be interpreted in multiple ways, associations help them to recognize an event (cf. Figure 1.1), a phenomenon that is primarily investigated in visual perception (Bar, 2004). For example, when parts of continuous speech are replaced with noise, people can still perceive the speech as being continuous. Furthermore, their interpretation of the distorted part of the speech depends on the meaning of the surrounding speech (phonemic restoration, Warren, 1970; Samuel, 1996).

1.3 Sound sources, events, and percepts

In the previous sections we have introduced how automatic systems and people recognize sound events. However, we have not yet clarified the relation between a source, an event, and the perception of an event. In contrast to vision, audition is by default not static; that is, something in the world has to happen or change to produce sound. Furthermore, a potential sound source has to be involved in some event to produce sound, and it can often produce multiple types of sound events. For example, a car can be parked, producing no sound, someone can accelerate it, producing a sound event that is caused by a process in the engine, or someone can stop the car, producing a different sound event that is caused by a different process. Sound event recognition describes the task of recognizing events, which are caused by a physical action involving a source.

Human perception of sound events is explained by several theories. The ecological approach to sound perception adopts the term ‘everyday listening’ to refer to listening to sound events in everyday life (as opposed to ‘musical listening’,

---

1 A mexican in a canoe.
Gaver, 1993). Its focus is on the invariant (constant in different situations) perception of the physics of an event. For example, a large group of studies has focused on the ability of people to hear some physical property in the sound produced by an object, such as the perception of object size (Carello et al., 1998; Kunkler-Peck and Turvey, 2000), or the ability to distinguish between bouncing or breaking objects (Warren and Verbrugge, 1984). However, ecological psychology is less concerned with the functional process of recognition involving the role of memory in perception. Because this aspect is essential for modeling real-world sound event recognition, our focus is more toward the information processing approach developed in cognitive psychology than on ecological psychology.

The information processing approach analyzes cognition and perception by abstract stages in the processing of a task (Anderson, 2005). Cognitive psychology is not concerned with the function of the brain. Hence, these abstract stages do not necessarily correspond to the processing stages in the brain. Theories about cognition are usually investigated with an experimental paradigm to confirm (or falsify) the hypothesized stages. This experimental approach was first empirically tested by Sternberg (1966) in a memory decision task. In a similar way, auditory perception can be analyzed as a succession of conceptual processing stages, from sensory transduction, via auditory grouping and categorization, to recognition (McAdams, 1993). A schematic overview of human perception of a sound event is depicted in Figure 1.4. When people perceive an object or event in the world, they match the grouped percept to a prototype or exemplar of a category (Rosch, 1975; Dubois, 2000)—categories are not represented in memory by membership conditions, but rather by the attributes of a prototype (Reed, 1972; Smith and Minda, 2000) or exemplar (Medin and Schaffer, 1978; Nosofsky and Zaki, 2002).1 Furthermore, this prototype triggers associations in memory, which provide predictions about properties of the environment, for example, what may be perceived next.

---

1 For example, when one tries to think of list of properties shared by all chairs, it will be quite impossible to come up with even one. For every property an exception can be found that would still be conceived of as a chair. Yet, when people see an object, they have no difficulties in determining whether or not it is a chair.


**Figure 1.4:** A schematic overview of human perception of a sound event. The perceived event is matched to a prototype or exemplar of an auditory event category, which triggers associations in memory. These associations provide predictions about properties of the environment (Bar, 2007).

### 1.4 Overview

The thesis is divided in two parts. In the first part we give fundamental background on recognizing sound events. Furthermore, we introduce a method that is based on a model of human memory to improve sound event recognition with knowledge of the context. In the second part we demonstrate the improved performance of two applications that integrate *signal-driven* methods with the context model proposed in the first part.

Real-world environments pose additional demands on sound event recognition over controlled conditions, such as an input of isolated sounds. Chapter 2 discusses how these demands can be managed by an automatic system for sound event recognition. In chapter 3 we review the research on human perception of sound events, with a focus on studies in cognitive science. Furthermore, we present an experiment to demonstrate how context can facilitate sound event recognition. This facilitatory effect is known in visual perception, but hardly investigated in auditory perception. Finally, in chapter 4, we introduce a model that incorporates context
into an automatic sound event recognition system, based on the findings in the
previous two chapters.

Although sound recognition in real-world environments has been used to dis-
tinguish between different types of sonic environments, such as parks and roads
(Aucouturier et al., 2007), automatic recognition of the sound events that consti-
tute a sonic environment is a new area of research with important applications that
require a different approach. In chapter 5 we present two experiments that show
the possibility to automatically recognize sound events in an unconstrained envi-
ronment with the combination of techniques for sound event segregation (Krijn-
ders, 2010) and the context model. Moreover, the context model is not limited to
audio input, but can be applied to input signals from other modalities as well, as is
demonstrated in chapter 6. In this chapter we apply the context model to improve
robot localization by disambiguating visual observations of a mobile robot.